CONSONANT CLASSIFICATION BASED ON TONGUE TIP TRAJECTORIES

ROJIN MAJD ZARRINGHALAM

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

MASTER OF SCIENCE

GRADUATE PROGRAM IN DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO

DECEMBER 2017

Abstract

In this thesis, I investigate an issue that is foundational to the development of a new class of novel game-based speech therapies. Whereas several prior computer-based approaches have focused on the use of clinical objectives that concern spatialized aspects of the tongue-tip trajectory (e.g., the targeting of improved accuracy in lingual-palate contact for certain phonemic segments), this line of inquiry focuses on the potential use of attributes relating to the speed and velocity of the tongue-tip trajectory as an alternative clinical objective. I situate my work in the body of prior work on the velocity characteristics of different phonemic segments. For speed and velocity-based clinical targets to be viable, however, it is necessary to characterize and to analyze the relative amounts of variability among and within talkers and phonemic segments with respect to speed-related characteristics. I describe our approach and the results of an analysis which focuses on a large kinematic speech dataset that includes multiple repetitions of 8 different phonemic segments (/t/, /d/, /k/, /g/) as plosives, (/s/, /sh/, /z/) as fricatives and (/tch/) as affricate by 17 talkers. Last, we provide an illustration of how such 'normative' (albeit speaker-dependent) speed and velocity profiles would be instantiated via an interactive scenario that could be included within an extant computer-based speech therapy system. I will repre-

sented the classification accuracy results of kinematic data using HMM and SVM classification techniques.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

The development of systems capable of providing assisted therapy is very important as a response to the societal challenge of providing health services. One of the objectives of Computer-Based Speech Therapy (CBST) is the development of an system that provides personalized therapy for speech disorders. In this thesis, I investigate an issue that is foundational to the development of a new class of novel game-based speech therapies. Whereas several prior computer-based approaches have focused on the use of clinical objectives that concern spatialized aspects of the tongue-tip trajectory (e.g., the targeting of improved accuracy in lingual-palate contact for certain phonemic segments), this line of inquiry focuses on the potential use of attributes relating to the speed and velocity of the tongue-tip trajectory as an alternative clinical objective. I situate my work in the body of prior work on the velocity characteristics of different phonemic segments. For speed and velocity-based clinical targets to be viable, however, it is necessary to characterize and to analyze the relative amounts of variability among and within talkers and

phonemic segments with respect to speed-related characteristics. The main goal of this study is to analyze the speed and velocity characteristics of a set of different consonant segments and to perform a feasibility study to determine the degree to which these characteristics can serve as the basis for on-line classification, as required for Computer-Based Speech Therapy (CBST).

1.2 Computer-Based Speech Therapy

Speech rehabilitation therapy refers to a wide range of services that are provided by Speech-Language Pathologists (SLP) for optimizing communication to increase an individual's quality of life. Current clinical practices in speech intervention and rehabilitation rely on a wide range of techniques, highly dependent on the conditions being treated and where the service is delivered, such as public and private practice clinics, hospitals, rehabilitation facilities, community clinics and academic departments.

Intelligibility is commonly used in assessment and is also used as a measure of progress during therapy. Assessment of intelligibility may be based on the visual and/or auditory perceptions of the client's speech or those of the family reported or self-reported improvements in communication. Intelligibility can also be assessed by a computer-based speech therapy (CBST) system, which can also provide helpful feedback (e.g., by displaying the correct position and shape of the tongue). CBST systems, in addition to acoustic signal inputs, also can make use of kinematic signal inputs. An example of this is a tongue-controlled computer game using the Tongue Drive System (TDS) for the rehabilitation of tongue motor function (Kothari 2014). These kinematic-based systems have the potential to provide new directions in motor speech

rehabilitation, especially in the case of neurodegenerative diseases, such as Parkinson disease, that have known speech motor symptoms.

Speech research is now taking advantage of Electromagnetic Articulography (EMA), a sensor technology that collects large volumes of real-time 3D data about the movement of the tongue and other articulators. An example is the WAVE system (Northern Digital, Canada). Tongue positions during the production of the lingual consonants can be measured using this point-parameterized electromagnetic tracking. Since most of the articulators are hidden from view and entail millisecond-duration movements, without such sensor technologies, these movements could not be studied easily. For computer-based speech therapy (CBST), the application of EMA represents great potential as a major step forward in clinical practice. This technology would allow CBST systems to use kinematic signals, to augment or to replace their current reliance on acoustic signals.

1.3 Research objectives

The objective of Computer-Based Speech Therapy (CBST) systems is to provide therapeutic outcomes via computationally-based means, via automatic signal acquisition, analysis, and feedback design.

A long-term research objective is to develop useful CBST systems that make use of kinematic signals for speech therapy interventions that concern the production of consonant segments. This type of CBST system is premised on the fact that each consonant segment has its own distinctive kinematic features, which serve as 'clinical targets' for recipients of speech therapy.

A ‘clinical target ’, in this context, is the objective of the clinical intervention, or the objective that the therapy is attempting to obtain. This thesis work will address the foundational issues, which is the degree to which each consonant segment has its own distinctive kinematic features. The investigation will concern each of eight different consonant segments, which cover three fricatives consonants (/s/, /z/, /sh/), four plosive consonants (/d/, /k/, /t/, /g/) and one affricate (/tch/).

1.3.1 Research Questions

My objective is to derive a classifier for eight different consonant segments and to determine the accuracy of classification. To address the research objective, I have structured this work around the following three questions:

1. Question 1: Given a talker’s own data concerning tongue-tip kinematic profiles for different consonant segments, how accurately can we classify the consonant segment of unknown tongue-tip kinematic profiles by that same talker?

2. Question 2: Given data representing the tongue-tip kinematic profiles for a pool of different talkers, how accurately can we classify the consonant segment of unknown tongue-tip kinematic profiles by a talker from outside that pool?

3. Question 3: Given data representing the tongue-tip kinematic profiles for a pool of different talkers, how accurately can we classify the consonant segment of unknown tongue-tip kinematic profiles by a talker from within that pool?

1.4 Approach

I outline the three questions of interest in this work and the suite of three studies that I designed in order to answer them. I will describe the techniques for analyzing and classifying kinematic speech data, including Procrustes distance, Support Vector Machine (SVM), and Hidden Markov Model (HMM).

I will derive the speed and velocity characteristics of the tongue trajectories for a set of 8 different consonants from a corpus of relevant data. I will perform a series of pre-processing steps, employ SVM and HMM as two classification approaches, and then analyze the results.

1.5 Thesis overview

Chapter 2 provides a literature review of speech motor control, Electromagnetic Articulography (EMA) technology, and Computer-Based Speech Therapy (CBST) systems and clinical targets. This review supports the need for classification-based approaches for kinematic speech data for clinical targets.

Chapter 3 describes the design space, the development process, and the requirements, and evaluation strategies. Three studies are described, each consisting of four tasks.

Chapter 4 presents the results of the studies and a contrastive analysis of the HMM and SVM classification approaches. A summary of the findings will be provided.

Chapter 5 presents a summary of the research project, identifies the key contributions and outputs of this project, and discusses future work.

Chapter 2

Background and literature review

2.1 Introduction

This chapter provides a summary of the speech science research literature that is related to the objectives of this thesis. The summary is structured as follows:

1. Overview of speech motor control
2. Overview of Electromagnetic Articulography (EMA) for the collection of kinematic speech data
3. Overview of Computer-Based Speech therapy (CBST) systems, with a focus on their clinical targets (i.e., the aspect of speech the CBST system is attempting to target therapeutically). Of particular focus is the degree to which prior CBST systems have employed aspects of tongue kinematics as the basis for training and feedback.
4. Overview of classification techniques for kinematic speech data

2.2 Speech Motor Control

In phonetics and phonology, articulation is the movement of the tongue, lips, jaw, and other speech organs (the articulators) in order to make speech sounds (Bauman-Waengler 2016).

Speech articulators are of two types: active and passive. During the articulation of sounds, the passive articulators, such as the upper lips, teeth, and hard palate, remain static. The active articulators, such as the tongue, move relative to the passive articulators. The tongue is generally regarded as the most important active articulator (Yunusova et al.).

Speech motor control is different from general motor control. It uses physical properties such as vocal tract limitations, aerodynamics and biomechanics in order to produce the relevant sounds (Fuchs and Perrier 2013).

An alveolar stop is a consonant sound which is made with the tongue in contact with the back of the teeth. Fricatives are consonants which are produced by forcing breath through a narrow opening which is made by positioning two articulators close together; these may be the upper lip against the lower teeth. The two different control strategies for each of alveolar stops and fricatives show significant differences in velocity, the amplitude of deceleration peaks, movement amplitude, and tongue tip movement (Fuchs et al. 2006).

A long consonant is a consonant that is held longer than the short consonants (Kloster Jensen 1968). Most languages, including English, do not have long consonants that are distinctive from short consonants. In Japanese, a language that does have distinctive short and long consonants, kinematic analysis of tongue movements showed that speakers decrease the speed of the tongue movement during the pronunciation of long consonants (Lofqvist 2010).

Dysarthria, a condition which negatively impacts the muscle control and coordination that is needed to produce speech, is characterized by unintelligible and slow speech (Sharp and Tasko 2011). It can be caused by neurological disorders that lead to tongue or throat muscle weakness, facial paralysis, brain tumors, brain injury and stroke. Speakers with dysarthria have a limited range of movements during alveolar consonant release (Kim et al. 2010).

The velocity of the movement of the speech articulators over time can be expressed as time-series data and then generalized as a function (or velocity profile). Changes in speaking rate can be observed in terms of changes in shape of this profile. The velocity profile changes from a symmetrical, single peaked function at fast speaking rates to an asymmetrical and multi-peaked profile at slow speaking rates (Adams et al.). This shape variation demonstrates the idea that alterations in speaking rate are associated with changes in motor control strategies. For example, the control strategy for speech gestures produced at fast and speaking rates consist of unitary movements, whereas the gestures produced at slow speaking rates include multiple sub movements (Adams et al.).

2.3 Electromagnetic Articulography

Electromagnetic Articulography (EMA) is a sensor technology that is based on tracking via electromagnetic induction. An electromagnetic field is produced by induction coils which are placed around the head, creating a current in any sensor coils placed within the field. Sensors are placed on the tongue and other non-stationary speech articulators that move during the speech and generate (x, y, z) coordinate data for each sensor position. For instance, the Wave Speech

Research System (NDI, Waterloo) produces time-stamped six-dimensional (6D) kinematic data within the generator fields: rotational and positional data in each of three dimensions. Another speech tracking technology is X-ray microbeam, which employs small pellets placed on the subject‘s tongue, teeth and nose. This type of tracking is accomplished by a very narrow x-ray beam passing through the subject area and detected by a sodium iodide crystal located behind the head. The dense pellets block the x-rays from reaching the crystal. The technique allows the study of speech patterns in real time (Westbury et al. 1990).

Of the two techniques, EMA provides a better alternative than X-ray microbeam for tracking the speech articulators. First, it uses low field-strength electromagnetic fields and thus reduces exposure to harmful radiation (Westbury et al. 1990). Second, it improves on early articulator tracking methods, such as x-ray microbeam, which were limited to two-dimensional (2D) tracking, and needed attention and detailed calibration (Perkell et al. 1992, Schönle et al. 1987). Subsequently-developed tracking methods provided full three-dimensional (3D) tracking of rotation and position (Kaburagi et al. 2005, Zierdt 1993). These methods of measurement became available via commercially available speech research measurement tools, such as the Wave Speech Research System (NDI, Waterloo) and the Carsten AG line of products (Carstens Medizinelektronik GmbH, Bovenden). The accuracy of these commercial products have been tested and shown to be sufficiently accurate for speech science research. One analysis of dynamic positional errors showed that 88% of them were < 0.5 mm and 98% of them were < 1.0 mm (Berry 2011, Kroos 2008, Yunusova et al. 2009). The tracking of the articulators, especially the tongue, using EMA tracking systems comes with challenges. These tracking systems require

that a sensor be placed at the point where data is to be collected. This means that sensors need to be affixed directly to the articulator, which can impact articulator kinematics and speech intelligibility (Katz 2006). Moreover, the placement of a tongue sensor is generally uncomfortable if it is closer than 1cm to the tip of the tongue (Hoole and Nguyen 1999, Perkell et al. 1992).

2.4 Computer-Based Speech Therapy (CBST) Systems

Clinical speech rehabilitation serves to create outcomes which are achieved by modifying speech parameters (Schröter-Morasch and Ziegler 2005). These speech parameters include *kinematic* parameters, such as articulator position or speed and *acoustic* parameters, such as pitch or volume.

There is a growing body of work focused on the applications of speech recognition and other computational techniques to derive results concerning the articulatory characteristics of speech, including kinematic and acoustic parameters, that can be applied in clinical rehabilitation contexts. The applications described in this review include those that focus on health and wellness, speech and language therapy, rehabilitation, and the assessment of treatment efficacy. The objective of Computer-Based Speech Therapy (CBST) systems is to provide therapeutic outcomes via computationally-based means, via automatic signal acquisition, analysis, and feedback design (Haworth 2016).

An example of a CBST system is a tongue-controlled computer game for the rehabilitation of tongue motor function (Kothari 2014). This project included the development of a Tongue Drive System (TDS) and two studies which investigated the effect of tongue disability, age and

sex on tongue motor performance following a tongue-training pattern using the TDS. The TDS allows individuals with disabilities to operate a computer, to control a powered wheelchair and to interact with their environments simply by moving their tongues. In one study, subjects with impaired tongue function and with dysarthria were matched with age and sex controls, all of whom participated in tongue training (study 1). In study 1, both pre- and post-intervention revealed that the tongue-disabled patients demonstrated relative poorer motor performance than healthy controls (p=0.005) overall with a significant effect of sex ($p < 0.05$). There was improvement in motor performance for both the tongue-disabled and healthy controls groups, and the difference between them was not significant. In the second study, healthy participants (both elderly and young) participated in tongue training (study 2). The study showed there were main effects of age ($p \leq 0.001$) on performance. Healthy young volunteers accomplished better motor outcomes than healthy elderly participants ($p \leq .001$). The studies provided evidence that age and degree of tongue disability has an effect on behavioral measures of tongue motor performance. TDS may be a new neurorehabilitation technique in treating tongue-disabled patients (Kothari 2014).

van Vuuren and Cherney (2014) described a animated virtual therapist (VT) application for delivering speech and language therapy to persons with aphasia (PWA). This work provides three different perspectives which focus on role, implementation and performance, respectively. The first perspective describes the typical roles and treatment that the VT performs in directing practice, participation and performance. The second perspective describes the modeling and implementation considerations for visible speech and tele-rehabilitation. The third perspective

concerns the analysis of the performance of the system for delivering language and speech therapy to people with aphasia. The system which they described can work across a number of different devices (e.g., as a application on a computer, mobile, or tablet device, or as a web application in a cloud-based client-server configuration which is suitable for tele-rehabilitation). The basis for therapy was oral production of scripts and short functional dialogs which were structured based on communication for everyday activities. Guidance and feedback are also provided interactively by the VT. The study showed that for persons with aphasia in a real-world setting, receiving treatment delivered by a VT can lead to faster learning.

In work by Katz et al. (2014) a virtual tongue and head model was developed. The model is animation-based, and the motion of the model is driven by the WAVE data acquisition system, which provides real-time information regarding the tongue and jaw movements of an instrumented subject during speech. The users of the system are able to see their tongue position in real time on a customized interface. This system provides real-time feedback for tongue targets related to place of articulation for American English consonants. The place of articulation of a consonant is the point of contact where an obstruction occurs in the vocal tract among an active articulator (typically some part of the tongue) and a passive location (typically some part of the roof of the mouth). The system uses spatial target zones corresponding to correct places of articulation for each talker. When the talker's tongue achieves the correct place of articulation for American English consonant, the talker receives augmented visual feedback. The system allows speakers to observe in real time how their tongue is moving as they utter speech sounds. Katz et al. (2014) investigated whether this system can accurately record talker's lingual place

of articulation for alveolar stimuli in a laboratory speech setting, and then provide real-time visual information concerning place of articulation which would be useful for speech training application. Preliminary data obtained for a group of adult talkers suggest this system can be used to reliably provide real-time feedback for American English consonant place of articulation targets, and that on-line visual feedback provided by the system may be used by talkers to improve accuracy for articulation during consonant production.

Yunusova et al. performed a study to determine the degree to which tongue position is unique among a set of lingual consonants (i.e., alveolar stops, such as /p/ and /k/, alveolar fricatives, such as /z/ and /s/, and postalveolar consonants, such as /sh/ and /ch/). Tongue positions during the production of these consonants were measured using point-parameterized electromagnetic tracking via the VAVE systems. Once point-parameterized methods for studying speech movements became available, positional targets that are reached by the tongue during speech have been defined in terms of ranges of acceptable positions (target regions) in two- or three-dimensional space (Guenther 1994). On the basis of the tongue positional data, the target region of the talker's tongue in terms of x, y, z tongue positions for each of consonants were extracted. Alveolar stops (such as /p/ and /K/), alveolar fricatives (such as /z/ and /s/), and the postalveolar consonants (such as /sh/ and /ch/) all displayed different target regions. The findings demonstrated that tongue positions are not unique for a talker completely. Cognates pairs are pairs of consonants which share the place (and manner) of their articulation, such as the pairs /d/ and /t/, /s/ and /z/, and /k/ and /g/. Voiced and voiceless cognates were found to share the location of their positional targets Postalveolar homorganic consonants, which are

consonant sounds that are articulated in the same position or place of articulation in the mouth (such as /m/ and /p/), were found to share the location of their target regions. The individuals characteristic of the palate and the speaking rate were the other variables which we found to be important in variation.

A virtual articulation teacher was developed by (Engwall 2012). It performed analysis of and provided feedback on phonetic features in pronunciation training, and showed the correct position and shape of the tongue via audiovisual feedback to demonstrate how different phonemes, with correct or incorrect pronunciation, can be distinguished in the articulatory space. The system made use of a cut-away display which is a graphical display of the head with certain parts removed in order to make the intra-oral articulations, such as jaw, tongue and palate, visible. By making parts of the face transparent, the system can show the correct position and shape of the tongue and provide audiovisual feedback on how to change incorrect articulations. For this system, a reverse kinematic approach was used. This approach entails the use of computational techniques which take acoustic input and attempt to derive the kinematics of the speech sound articulator from that. In an observation study Engwall (2012), seven subjects used the system, and articulatory changes were observed and were conjectured to be attributable to the audiovisual feedback that they received during training. The short-term changes in articulation observed in users was a positive sign that articulatory feedback instructions could be appropriate in computer-assisted pronunciation training.

2.5 Classification Techniques for Kinematic Speech Data

The articulatory distinctiveness of consonants and vowels based on the tongue movement has been previously investigated using classification methods in several prior research projects.

2.5.1 Classification Using Procrustes Distance

Procrustes analysis is a robust shape analysis technique which has been successfully applied for shape classification and object recognition (Goodall 1991). In Procrustes analysis, a shape is presented as a set of ordered landmarks on its surface. The Procrustes distance between two shapes is computed by first aligning the two shapes using their centroids. Then, both shapes are scaled to a unit size, and then one shape is rotated to match the other and the minimum sum of the Euclidean distances between their corresponding landmarks is obtained. Thus, the distance reflects shapes differences once the scaling, rotational and locational effects have been removed.

Procrustes distance has been previously applied specifically in speech signal analysis, when it was used to measure the articulatory distinctiveness of the movement of 19 sound segments: eight major English vowels and eleven English consonants (Wang et al. 2011). Procrustes analysis was designed for static shape analysis. Procrustes distance between vowel and consonant shapes defined by sampled tongue and lip motion paths was proposed as an index of the articulatory distinctiveness between vowels and consonants. The motion paths of six sensors, which were attached on the lips and the tongue, were collected for each subject, generating lip and tongue movement time-series data. This time-series data of sensor locations, which were recorded using EMA, went through a sequence of preprocessing steps prior to analysis. First,

the head movements were subtracted from the lip and tongue locations. Second, noise was removed by applying a low pass filter of 10 Hz. Third, all sequences were parsed to segments that corresponded to single speech sounds (the vowels and consonants). The segmentation was done manually by aligning the movement data with acoustic data recorded synchronously. When the participant spoke, the 3-D location data of the sensors were recorded and saved as x,y and z axes. Here, x, y, and z are defined as spatial dimensions width (left-right), height (up-down) and length (front-back) in a 3-D coordinate system. Since the movement along the x axis is not considered significant in normal speech production, only the y and z coordinates of the sensors (i.e., upper lip (UL), lower lip (LL), T1 (Tongue Tip), T2 (Tongue Body Front), T3 (Tongue Body Back) and T4 (Tongue Root)) were used for analysis. The motion path trajectories of all the six sensors for each vowel and consonant were down-sampled to 10 landmarks (Wang et al. 2011).

A Procrusters-based classifier was developed using three steps: (1) representative shape derivation; (2) articulatory distinctiveness derivation; and (3) derivation of classification results.

For the first step, a representative shape was derived for each per-subject sound segment. The average shape for each phoneme was determined via the averaged coordinates of corresponding landmarks of all samples for that phoneme. The average shapes of all samples for each phoneme were calculated based on the average positions of corresponding landmarks of all samples for the vowel and consonant. The average shapes were also used as references for the phoneme, there is one average shape for each vowel and consonant. Through this process, the motion paths for a given vowel or consonants were spatially integrated as a composite shape, which created a representative shape for that vowel or consonant.

For the second step, for the representative shapes for each of the 11 vowels and consonants, the Procrustes distances between all pairwise combinations were calculated. Distance (distinctiveness) matrices were used to generate articulatory vowel and consonant spaces using multi-dimensional scaling.

For the third step, recognition of a particular sound segments was made on the basis of the shortest distance between it and the possible average shapes. This Procrustes-based classifier resulted in an average classification accuracy for vowels of 91.7% and for consonants of 91.37%.

Although Procrustes-based approaches can be used for classification, over the last decade, support vector machines (SVMs) have become the reference for many classification problems because of their flexibility, computational efficiency and capacity to handle high dimensional data (Nguyen and De la Torre 2010). SVMs have the potential to exploit more information about kinematic signals than the Procrustes-based approach. SVM have become a popular tool in time series forecasting due to their remarkable characteristics, such as good generalization performance, the absence of local minima, and the sparse representation of solution (Cao and Chong 2002).

2.5.2 Classification Using Support Vector Machines (SVMs)

Support vector machines (SVMs) are supervised learning models which are related to learning algorithms. Supervised learning is the machine learning task of deducting a function from labeled training data (Mohri et al. 2012). Given a set of training examples, each marked with a category label, a SVM training algorithm builds a model that can assign new examples to one

of its categories.

The Maximal-Margin Classifier is a hypothetical classifier that best explains how SVMs work. Numeric input variables form an n-dimensional space. For two input variables (X_1 and X_2), this would form a two-dimensional space. A hyperplane is constructed that splits the input variable space, selected to best separate the points in the input variable space by their class. In our two-dimensional example, this would be either class 0 or class 1; the hyperplane would be a line; and the following expression shows if all of the input points can be completely separated by this line:

$B_0 + (B_1 * X_1) + (B_2 * X_2) = 0$

The coefficients B_1 and B_2, which determine the slope of the line, and B_0, which is the intercept, would be found by the learning algorithm. By plugging in input values into the line equation, one can calculate the position of a new point relative to the line, thus determining its class. Above the line, the equation returns a value greater than 0 and the point belongs to the first class (class 0). Below the line, the equation returns a value less than 0 and the point belongs to the second class (class 1). A value close to the line returns a value close to zero and the point may be difficult to classify. If the magnitude of the value is large, the model may have more confidence in the prediction. The distance between the line and the closest data points is referred to as the margin. The best or optimal line that can separate the two classes is the line that as the largest margin. This is called the Maximal-Margin hyperplane. The margin is calculated as the perpendicular distance from the line to only the closest points. Only these points are relevant in defining the line and in the construction of the classifier. These points are

called the support vectors. They support or define the hyperplane.

The hyperplane is learned from training data using an optimization procedure that maximizes the margin.

Support Vector Machines are based on the concept of decision planes that define decision boundaries (Hill and Lewicki 2007). A decision plane is one that separates between a set of objects having different class memberships. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap. New examples, based on which side of the gap they fall, are mapped into that same space and predicted to belong to a category. The learning of the hyperplane in linear SVM is done by transforming the problem using linear algebra. A powerful insight is that the linear SVM can be rephrased using the inner product of any two given observations, rather than the observations themselves. The inner product between two vectors is the sum of the multiplication of each pair of input values. For example, the inner product of the vectors [2, 3] and [5, 6] is 2*5 + 3*6 or 28.

In addition to performing linear classification, SVMs can perform a non-linear classification using what is called the kernel trick, by mapping their inputs into high-dimensional feature spaces. In general, it depends on dataset characteristics, the numbers, number of samples and inter-relationship among input variables.

For SVMs like many other supervised learning problems, feature selection is important for a several reasons. First, feature extraction involves reducing the amount of resources required to describe a large set of data. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of vari-

ables generally requires a large amount of memory and computation power, also it may cause a classification algorithm to overfit to training samples and generalize poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems, while still describing the data with sufficient accuracy (Alpaydin 2014).

A study conducted by Wang et al. (2009) investigated the classification accuracies obtained via three techniques — Support Vector Machines (SVMs), Neural Networks, and Decision Trees — for recognizing vowels from articulatory position time-series data. This approach directly mapped articulatory position time-series data to vowels without extracting articulatory features such as mouth opening. A single-speaker dataset of eight major English vowels acquired using Electromagnetic Articulography (EMA) was used. The position data of each phoneme was time-normalized and sampled to fixed widths. The width was fixed because all classifiers require that input vectors of attributes have the same size. The study demonstrated that the SVM accuracies were higher than the other techniques for recognizing speech from articulatory movements.

In another series of studies conducted by Wang et al. (2012), classification was performed using Support Vector Machines (as opposed to the Procrustes-distance based technique that described previously, in section 2.5.1). In this study, the time-series articulatory position data was used as the features for the SVM classifier (only tongue and lip movement data was used, and did not include additional articulatory features such as mouth opening). A set of time-series data (pre-segment articulatory movement data) was sampled to fixed-width vectors of articulatory positions of the tongue and lips and time-normalized. Acoustic data was also obtained

during data collection and was used only for segmenting the kinematic data (and not used for recognition). Thus, there is no computational processing required to calculate the correlation between input data and target sounds in the prediction stage or to extract articulatory features.

Once trained, the SVM classifier was used to recognize words from the testing data set of tongue and lip movement data. To perform this recognition, the data was used in its continuous and unsegmented form (as opposed to the segmented form used for training); a prediction window, with variable boundaries, was used on the testing data to identify segments for classification. Candidates segments were identified when the segment in the prediction window (represented by its left and right boundaries, w_l and w_r) produced a probability value that exceeded the candidate threshold (which was obtained empirically from training data). The segment, once identified in this way, was then classified according to the highest probability value. Thus, window boundaries were derived to recognize words and their locations within the window based on the probabilities which were returned by SVM by providing probability estimates transformed from SVM decision values. Thus, segmentation and identification was conducted together.

All possible word lengths (within the length range of training words with a step size) were considered and the maximum probability was returned as the probability for a time point. The algorithm is based on the premise that a word has its highest matching probability given an observation window with an appropriate starting point and width. A trained classifier that derives these matching probabilities is embedded into the algorithm. As a result, the algorithm missed 1.93 words in a sequence of twenty-five words, with, an average latency of 0.79 seconds

for each word prediction, using a data set of 5,500 isolated word samples collected from ten speakers. The results demonstrate the effectiveness of the approach and its potential for building a real-time articulation-based silent speech interface for health applications.

For the SVM approach-based classifier approaches described by Wang et al. (2013), vowel and consonants classification was performed to quantify the articulatory distinctiveness of 8 major English vowels and 11 English consonants based on tongue and lip movement time series data using a support vector machine. Vowel classification accuracies of 89.05% and consonant classification accuracies of 88.94% were obtained which is lower than obtained results of procrustes based approach with 91.7% and 91.37% for vowels and consonants, respectively (Wang et al. 2011). In this study they have extended the typical use of Procrustes analysis, which was designed to analyze static shapes (i.e., shapes that do not change over time), to the analysis of time-varying shapes (i.e., shapes that change overtime).

Speech recognition software that depends on knowledge of the speaker's particular voice characteristics is called speaker dependent. Speaker independent systems are able to recognize the speech from different users by placing limitations on the contexts of the speech (the words and phrases). Wang et al. (2014) reported an across-speaker articulatory normalization approach based on Procrustes matching. A dataset of short functional sentences was collected from seven English talkers. A support vector machine was then trained to classify sentences based on normalized tongue and lip movements. The accuracy of word recognition for speaker-dependent speech recognition was significantly higher than speaker-independent without normalization ($p < 0.01$) and the recognition accuracy for speaker-independent recognition with the normaliza-

tion approach was also significantly higher ($p < 0.01$) than for the speaker-independent without normalization approach. The reason for selecting a subset of features from data set is for preserving or improving the discriminative ability of a classifier.

In addition to SVM-based approaches for classification, other techniques, such as Hidden Markov Models (HMM) have also become prevalent for signal data. HMM-based classification has the potential to exploit more information about the kinematic signal than the Procrustes-based approach and are typically expected to perform better than Procrustes-based classifiers.

2.5.3 Classification Using Hidden Markov Model (HMMs)

Application areas for HMMs include speech recognition, gesture recognition, language modeling, stock price prediction and many more. HMMs are also used for detecting non-language speech sounds in a speech or an audio signal, to improve the performance of speech processing and to enhance classification applications (Tomaschek et al. 2013). The ability to detect different classes of sounds could be considered as a pre-processing step and may significantly improve the performance of speech processing applications (Richardson et al. 2003).

2.5.3.1 Dynamic Time Warping

Dynamic time warping (DTW) is a method for determining the similarity between two time-based data series in a way that accounts for differences in time duration. The method calculates an optimal match between two time-based data series, with certain limitations (continuity constraints, restriction windows, endpoints, local distance definitions, and other limitations). The

algorithm calculates the local compression and/or the local stretch to apply to the time axes of one time-series data set (query) in order to map it onto the other optimally (Giorgino 2009). Bossemeyer et al. (1988) described a DTW-based algorithm approach to the problem of finding keywords in audio data to recognize vocabulary words in the context of unconstrained speech. The technique entails matching a keyword template, which is one of five defined vocabulary words (i.e., collect, calling card, person, third number, and operator), to the unknown speech at each starting frame of an utterance by considering voicing duration and energy level. This algorithm had 90% recognition accuracy rates on utterances in unconstrained speech for vocabulary words spoken in unconstrained speech are being achieved.

Although DTW approaches can be used for classification, other techniques, such as Hidden Markov Models (HMM) have become more prevalent. HMM-based classification are thought to exploit more information about the signal than the DTW-based approach (Bossemeyer et al. 1988), and HMM classification results are typically better than DTW-based results using same data (Wilpon et al. 1990).

2.5.3.2 HMM classification

A hidden Markov model (HMM) is a statistical Markov model of a sequence of feature vector observations. The system being modeled is assumed to be a Markov process (memoryless process) with hidden (unobserved) states. A sequence model or sequence classifier is a model whose job is to assign a label (or class) to each unit in a sequence, thus mapping a sequence of observations to a sequence of labels.

A HMM is a probabilistic sequence model: given a sequence of units (letters, words, sentences), it computes a probability distribution over possible sequences of labels for the inputted units and chooses the best label sequence. A HMM has a set of states, each of which has limited number of transitions and emissions, where each transition between states has an assigned probability. Each model starts from a start state and ends in an end state.

Classification is fitting the emission probabilities for each state based on observed data and corresponding class. The current state of the model depends only on the previous state. In a simple Markov models (like a Markov chain), the transitions depends on the current state and the transition probability matrix, and the state is directly visible to the observer. Therefore, the state transition probabilities are the only parameters. The goal is to make a sequence of decisions where a particular decision may be influenced by earlier decisions. In a HMM, the state is not directly visible, and the Markov process is hidden, but the output, dependent on the state, is visible. Each token in a sequence is assigned a label. Labels of tokens are dependent on the labels of other tokens in the sequence, particularly their neighbors. Each state has a probability distribution over the possible output tokens. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states. The term 'hidden' does not refer to the parameters of the model, but rather it refers to the state sequence through which the model passes; the model is still referred to as a 'hidden' Markov model even if these parameters are known exactly (Rabiner 1989).

Hidden Markov models (HMMs) provide an effective and simple framework for modeling time-varying vector sequences. The use of a HMM serves to systematically explore the solution

space, by looking at a different number of states.

2.5.4 HMM Description

A hidden Markov model (HMM) can be described as follows:

1. Hidden states Q = q_i , i = 1, . . . , N.

2. Transition probabilities A = a_{ij}= P(q_j at t +1 — q_i at t), where P(a | b) is the conditional probability of a given b, t = 1, . . . , t is time, and q_i in Q. A is the probability that the next state is q_j given that the current state is q_i.

3. Observations (symbols) O = o_k , k = 1, . . . , M .

4. Emission probabilities B = b_{ik} = $b_i(o_k)$ = P(o_k | q_i) , where o_k in O. B is the probability that the output is o_k given that the current state is q_i.

5. Initial state probabilities II = p_i = P(q_i at t = 1).

The model is characterized by the complete set of parameters: I = A, B, II .

As an example, assume there are two classes A and B; a HMM can classify an unknown sequence s to one of the A or B classes.

2.5.4.1 Training and Testing the HMM

In order to train a HMM for classification, the data set is split into training and testing parts. The standard approach is to separate the data sets into one data set for each class. A HMM is

trained per some specific percentage of the data set and then the test set is constructed from the reminder. A HMM can be trained on a given observation sequences, mapped on a per-sequence to one of the classes (Stamp 2015). Training the data that belongs to one class can be done separately, so that one HMM per class is trained. The standard algorithm for HMM training is the forward-backward algorithm. The algorithm trains both the transition probabilities A and the emission probabilities B of the HMM. There are HMM packages in Matlab and in Weka, which is a open source machine learning software. Once training is complete, then testing is done. This is performed by checking each trained model and determining which model produces the observed data with the highest likelihood.

2.5.5 HMM Modeling for Speech

HMMs are one of the most successful technique used in automatic speech recognition (ASR) systems. Automatic Speech Recognition (ASR) is a technique to convert the acoustic (or kinematic) signals of speech into words. The recognized words could be an input for a natural language processing or a final output (Richardson et al. 2003).

To recognize an utterance, the probability metric according to each model is computed and the model with the best fit to the utterance is chosen. The first step in building the model is to identify the components of the audio signal that are good for identifying the linguistic content and also, for the sake of exclusion, other components which carries non-relevant information like background noise, emotion etc. The identified components are then extracted as features.

A frequently used feature when analyzing acoustic signals is the Mel Frequency Cepstral

Coefficient (MFCC) (Hossan et al. 2010). It is a leading approach for speech feature extraction, the basis for identifying the components of the audio signal that are good for identifying the linguistic content and for discarding other components which carry information like background noise, emotion etc. The job of MFCCs is to accurately represent these features. In work by Fajardo and Kim (2014), the MFCC were derived from recorded speech that included Filipino vowels and consonants phonemes data and were used to train a prototype HMM. Filipino vowel phonemes were recognized with 90.8% accuracy for independent speakers and with 94.5% accuracy for dependent speakers. Since the dependent speakers were not common to the training data, it is expected that the accuracy results acquired from the independent speakers would be lower than the accuracy results obtained from the dependent speakers.

2.5.6 HMM-DNN modeling classification

A deep neural network (DNN) is an artificial neural network (ANN) with multiple hidden layers of units between the input and output layers. The DNN-HMM takes advantage of DNNfis strong representation learning power and HMMfis sequential modeling ability to speaker-independent silent speech recognition. This technique was recently used in to perform speaker-independent silent speech recognition from tongue and lip movement data (Wang and Hahm 2015).

2.6 Conclusion

Section 1.1 of this chapter presented the objectives of this chapter. Section 1.2 provided a summary of speech motor control and differences in motor control strategy for different speech

sounds. Articulatory characteristics were reviewed and the applications of speech recognition obtained results clinically were reported. In section 1.3, the use of Electromagnetic Articulography (EMA) was discussed. EMA provides the possibility of tracking the speech sound articulators. Section 1.4 presented the computational techniques which are involved in Computer-Based Speech Therapy (CBST) systems, with a focus on those which make use of articulatory clinical targets. Additionally, some techniques that have been used in the analysis of acoustic speech data was discussed. Section 1.5 provided the classification techniques for kinematic speech data including the Procrustes, SVM and HMM-based approaches. Previous use of the Procrustes analysis for shape classification and object recognition were discussed. SVM as a classification technique for regression and classification analysis was characterized, and different works which have used SVM for recognizing speech from articulatory movements were reviewed. Dynamic time warping (DTW) was discussed. The Hidden Markov Model (HMM) was described, and the training and testing of the HMM and HMM modeling for speech was reviewed. Previous use of HMM classification techniques, one of most successful classification techniques for speech data, was discussed. The classification accuracy obtained from different application domains using these techniques was summarized.

Chapter 3

Study Design

3.1 Introduction

In this chapter, I outline the three questions of interest in this work and the suite of studies that I designed in order to answer them. All of the questions below refer to *tongue tip kinematic profiles*, which refers to the time-series three-dimensional movement data about the path of the tip of the tongue during the articulation of a speech. In particular, I will be examining the particular consonant segments of /t/, /d/, /k/, /g/, /tch/, /s/, /z/, and /sh/. I will describe how I derive velocity- and speed-related features from the kinematic profiles of these speech sounds.

3.2 Objectives

1. Question 1: Given a talker 's own data concerning tongue-tip kinematic profiles for different consonant segments, how accurately can we classify the consonant segment of unknown tongue-tip kinematic profiles by that same talker?

2. Question 2: Given data representing the tongue-tip kinematic profiles for a pool of different talkers, how accurately can we classify the consonant segment of unknown tongue-tip kinematic profiles by a talker from outside that pool?

3. Question 3: Given data representing the tongue-tip kinematic profiles for a pool of different talkers, how accurately can we classify the consonant segment of unknown tongue-tip kinematic profiles by a talker from within that pool?

3.3 Approach

I now describe the design of three studies, one study to derive the answer for each question. Each study requires preparation of its own data sets, all of which will be derived from the main data set, which is described first.

3.3.1 Data Collection

3.3.1.1 Stimuli

Kinematic data corresponding to tongue-tip trajectories for eight lingual consonants segments was collected under a study protocol approved by the University Health Network Research Ethics Board (Certificate: 13-6235-DE Visual Feedback Systems in Speech Rehabilitation). This stimuli set is described in detail by Rudy (2011) and an overview is provided here.

The set of eight consonants segments consists of /t/, /d/, /k/, /g/, /s/, /z/, /tch/, /sh/, which cover three different *manners* of production and three different *places* of production:

The three different manners of production include:

i. plosives: /t/ /d/ /k/ /g/,

ʌ. affricates: /tch/(‘tch’ is also written as /t∫/ and is the ‘ch’ in ‘leach’),

ɜ. fricatives: /s/, /z/, and /sh/ (‘sh’ is also written as /∫/ and is the ‘sh’ as in ‘ship’)

Plosives are the kinds of sounds — associated with the letters p, t, k, b, d, and g — in which air flow from the lungs is interrupted by a complete closure being made in the mouth. Fricatives — associated with the letters s, z, ‘sh’ as in ‘ship’, and the ‘s’ in ‘vision’ — are characterized by a fihissingfi sound which is produced by the air escaping through a small passage in the mouth. Affricates — associated with the letters ‘ch’ in ‘leach’ and the ‘j’ in ‘jump’ — begin as a plosive and end as a fricative.

The three different *places* of production include:

1. post-alveolar: /sh/ /tch/,

2. alveolar: /t/ /d/ /s/ /z/, and

3. velar: /k/ /g/

In articulatory phonetics, *place* of articulation refers to the point of contact where an obstruction occurs in the vocal tract between an articulatory gesture, an active articulator (typically some part of the tongue), and a passive location (typically some part of the roof of the mouth). The place of production for the lingual consonants of focus in this study are summarized in table 3.1.

In this data collection protocol, each consonant segment is placed with a vowel before and after, to create a vowel-consonant-vowel (VCV) segment. Three “corner” vowels were employed:

		place of articulation		
		alveolar	post-alveolar	velar
manner	plosive	t d		k g
	affricate		tʃ dʒ	
	fricative	s z	ʃ ʒ	

Table 3.1: Manner and place of production for the eight lingual consonants /s/, /tch/, /z/, /sh/, /t/, /g/, /k/,/d/

/a/, /i/, /u/, to create a set of 24 different VCVs (3 vowels combined with 8 consonants). Theses carrier vowels were selected to increase the space where consonants are generated in the vocal tract. For instance, for the sound /d/, the 3 VCVs would be 'ada' , 'idi' and 'udu'. Each VCV is embedded in a carrier phrase ("It's .... game") to produce a set of 24 distinct stimuli items. The phrase was chosen in consideration of co-articulatory influences, to simplify the identification of acoustic segment boundaries in post-processing.

3.3.1.2 Subjects and Procedure

Subjects were recruited on the University of Toronto, St. George Campus. 17 healthy adult speakers with no history of speech disorders participated in this study (10 male, 7 female). The average age of the female group was 28.4 (SD = 6.1, Range = 25-43) and the average age of the male group was 32.3 (SD = 8.5, Range = 25-49). All participants were native speakers of Canadian English. Five speakers were from different parts of Western Canada and the remaining twelve

speakers were from Eastern Canada. Each participant was assigned a unique identifier, such as "W02". Participants were asked to repeat each stimulus item 8 times at a comfortable loudness and speaking rate.

3.3.1.3 Data Acquisition

The WAVE electromagnetic system was used for tracking the movement of speech articulators (NDI, Waterloo, ON, Canada). The system samples the movements of six sensors that are affixed to the speech sound articulators and sampled within a generated electromagnetic field at 100 Hz in three dimensions and logs the location into session files. The sensors are 2mm diameter and are illustrated in Figure 3.1a.

Two sensors were glued using PeriAcryl Oral Tissue Adhesive, a non-toxic dental surgical glue to the mid-sagittal surface of the tongue blade (TB) and tongue dorsum (TD). Figure 3.1b illustrates the approximate placement of these two sensors, in locations 'TB' (tongue blade) and 'TD' (tongue dorsum). The TB sensor was placed approximately 1 cm from the tip of the tongue, and the TD sensor was glued approximately 2 cm behind TB. The positions of the sensors were measured using a ruler and recorded for each speaker. Sensors were recorded at the sampling rate of 100 Hz. Acoustic signals were acquired simultaneously with speech movements at 22 KHz, using a professional lapel microphone (Countryman B3P4FF05B) positioned approximately 15 cm away from the mouth.

The repetitions for a single stimulus item (each VCV phrase "It's .... game") were recorded into a two files: a sound file and a tab-delimited text file consisting of (x, y, z) coordinates for

Figure 3.1: The WAVE sensor system (left) and the sensors on a speaker's tongue (right)

each sensor. For each stimulus sentence, further post-processing was performed by speech lab technicians. Kinematic signals were low-pass filtered at 15 Hz using a zero phase digital filter (8-pole Butterworth) (Yunusova et al.). The data was then segmented into individual repetitions using a manual process and the VCV segment extracted. Recordings were screened by a listener to ensure that only correctly produced sentences were included for analysis. Segments with tracking errors were excluded as well (less than 1% of data contained such errors). The average speaking rate is 273.37 millisecond per syllable, with the maximum at 360.77 and minimum at 219.71. The data was resampled at uniform 10 msec intervals.

3.3.1.4 Uncertainly in the Data

Recordings were screened by a listener to ensure that only correctly produced sentences were included for analysis. Segments with tracking errors were excluded as well (less than 1% of data contained such errors). The positions of the sensors were measured using a ruler and recorded for each speaker. The TB sensor was placed approximately 1 cm from the tip of the tongue, and the TD sensor was glued approximately 2 cm behind TB. Because of the challenges in measuring

the tongue and affixing sensors precisely, it is possible that the target position of the sensor and the actual position will differ slightly (on the order of millimeters), which may contribute as a source of error in the data. As well, the tracking instrument itself contributes a source of error. 88% of them were < 0.5 mm and 98% of them were < 1.0 mm (Berry 2011, Kroos 2008, Yunusova et al. 2009).

3.3.1.5 Tongue-tip Kinematic Profiles

In this study, the movement of the tongue is the focus. Data from the sensor on the tip of the tongue blade (TB) will be used and the data from the back (dorsum) of the tongue (TD) is not used.

Speed is the rate of change of distance with time and is expressed as distance moved (d) per unit of time (t), ignoring direction. Speed values were derived from each pair of sequential sensor coordinate positions in the data files by computing the length of the path connecting the two positions (corresponding to the speed of movement over a 10 msec duration). From these values, a speed profile of the tongue tip for each of VCV segment was derived. A sample of speed profiles for talker W02 for the consonant segment 'ASA' is shown in Figure 3.2.

3.3.1.6 Dynamic Time Warping

Dynamic time warping (DTW) was performed on the both speed and velocity profile time series data. Dynamic time warping (DTW) is a method that calculates an optimal match between two time series with certain limitations. The algorithm calculates the required compression or the

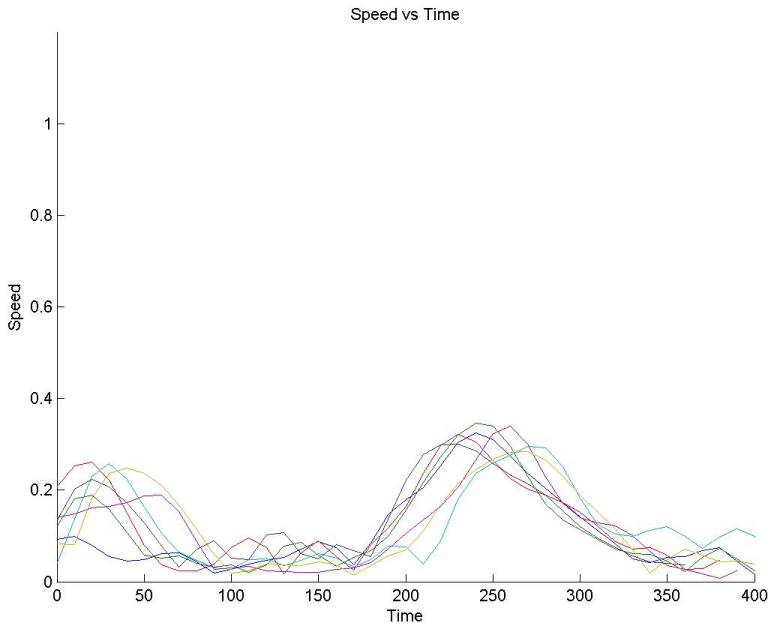

Figure 3.2: Speed curves before applying dynamic time warping, 8 repetition of ASA by subject2

local stretch to apply to the time axes of two time series data sets in order to map one (query) onto the other (reference) optimally (Giorgino 2009).

Using this technique, the speed and velocity profiles of each VCV segment were time-aligned on a per-speaker and per-VCV basis, using the first VCV repetition of the speaker as the reference. A script was prepared to implement this technique using the R programming language and software environment. The function `dlistWarped` from `dtw` library was used in the script. After the DTW processing step, all repetitions of a given VCV segment by a given talker will have been time-matched. A sample of speed profiles for talker W02 for the consonant segment 'ASAfi after applying dynamic time warping is shown in Figure 3.3.

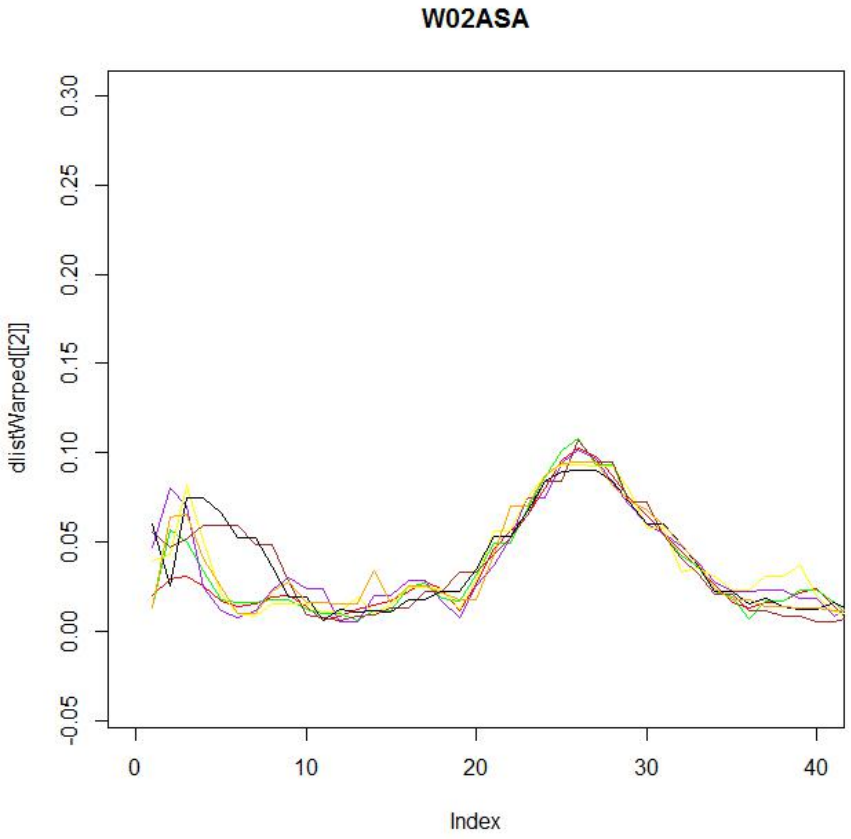

Figure 3.3: Speed curves after applying dynamic time warping, 8 repetition of ASA by subject2

3.4 Velocity-based Feature Derivation

Velocity profiles were derived for the kinematic time-series data of all repetitions produced by all 17 speakers for all of the stimuli items. Velocity is a vector quantity which is calculated with respect to direction. It is the measurement of change in rate and direction of an object. The initial data set contained points of the x, y and z coordinates of the location of the tongue tip for every sample point from pronouncing a given VCV phrase. Using Matlab functions, velocity vectors were calculated for every adjacent coordinate pair in the kinematic time-series data sets, resulting in a secondary dataset containing the velocity vectors resulting in sequences of data.

3.5 Speed-based Feature Derivation

The speed and velocity profiles that were prepared were then used the basis for feature derivation. For each of the three research questions, I intend to employ the SVM and HMM classification techniques. Below I describe the methods used to extract these features employed for each of these techniques.

3.5.1 Supporting Data Exploration

As a first step, all of the speed profiles were examined manually. To support this, an interactive info vis app was constructed using the Shiny library Chang et al. (2015), which is an add-on to the R programming environment. The app allows the user to select among VCV, talkers, and repetitions in order to explore the data set. For this app, the W* subject identifiers were

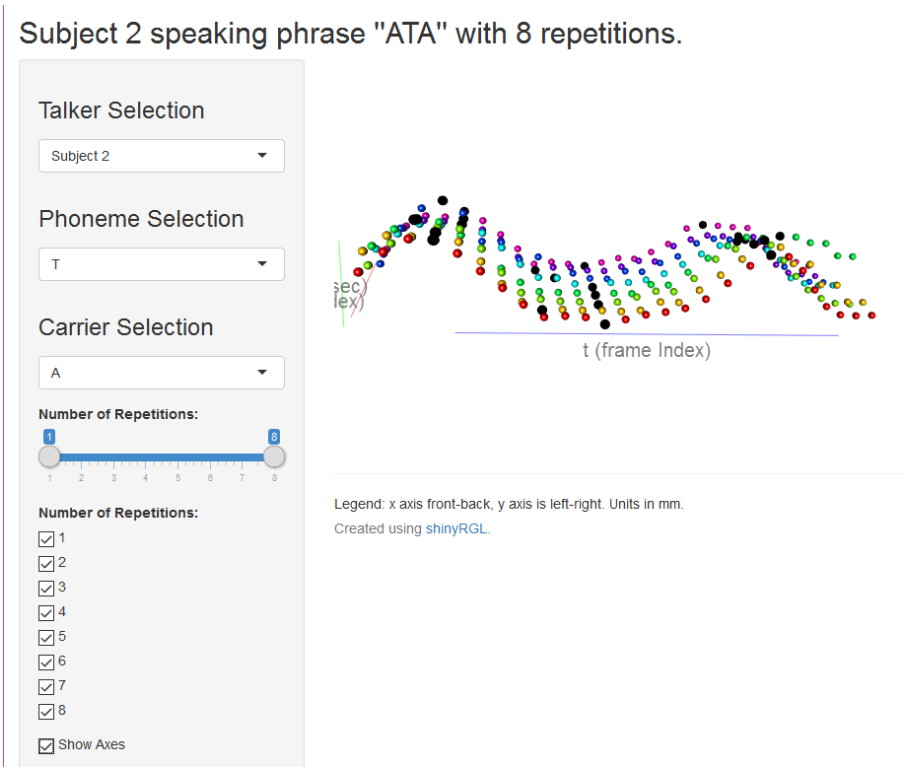

Figure 3.4: Data for 8 repetition of ATA by subject 2, visualized using the data exploration app.

remapped to the index values 1…17. A screen shot in shown is Figure 3.4.

3.5.2 Feature Set 1

Manual inspection of the data revealed that practically all speed profiles demonstrated a ”U” shape, with a core sequence of deceleration-acceleration phases for the tongue tip. An example is shown in Figure 3.5.

A set of six features was extracted from each speed profile. These features are drawn directly

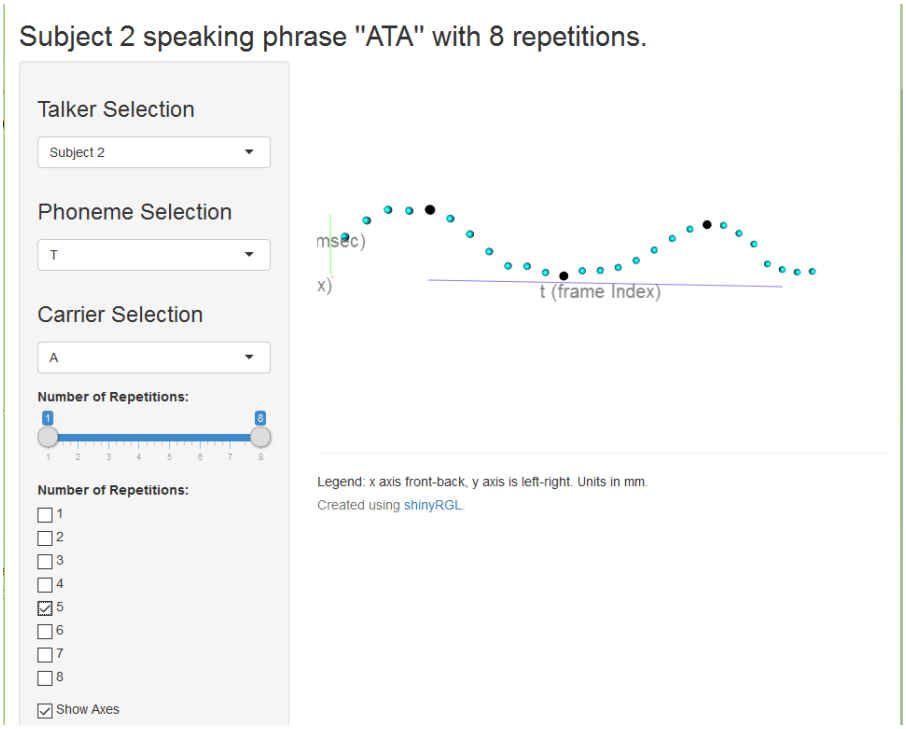

Figure 3.5: Data for one repetition of ATA by subject 2, visualized using the data exploration app.

from the three polynomial coefficients of two polynomial curves that were fit to the speed profile. One polynomial was fit to deceleration phase and the second polynomial was fit to the acceleration phase.

This procedure is contingent on first locating the two phases. The conjoined phases correspond to the segment that occurs between the two general maxima points. The minimum point in between marks the transition from deceleration to acceleration.

To proceed, I first located the transition point. To do so, I located the set of maxima within the speed profile. In many cases, there were more than two local maxima. I then located the first and the last local maxima point in the set of maxima. Then I found the absolute minimum point between these two local maxima. After finding that absolute minimum point, I located the absolute maximum before the minimum point and also another maximum point after this minimum. These three derived points are used to locate the beginning and end of the deceleration and acceleration segments respectively.

Using Matlab functions, the first curve, a polynomial of degree 2, $y = a_1x^2 + b_1x + c_1$, was fit to the deceleration segment. This segment consists of the series of data points that occur between the first absolute maximum and the absolute minimum. The minimum of the parabola aligned with minimum of the fitted polynomial. The coefficients of the parabola were derived, $(a1, b1, c1)$. Then the same procedure was done to the acceleration segment. This segment consists of the data points between the absolute minimum and second absolute maximum. The minimum of the parabola was aligned with minimum of the fitted polynomial. The set of coefficients for the second parabola were derived, (a_2, b_2, c_2). The result is a total of 6 features for

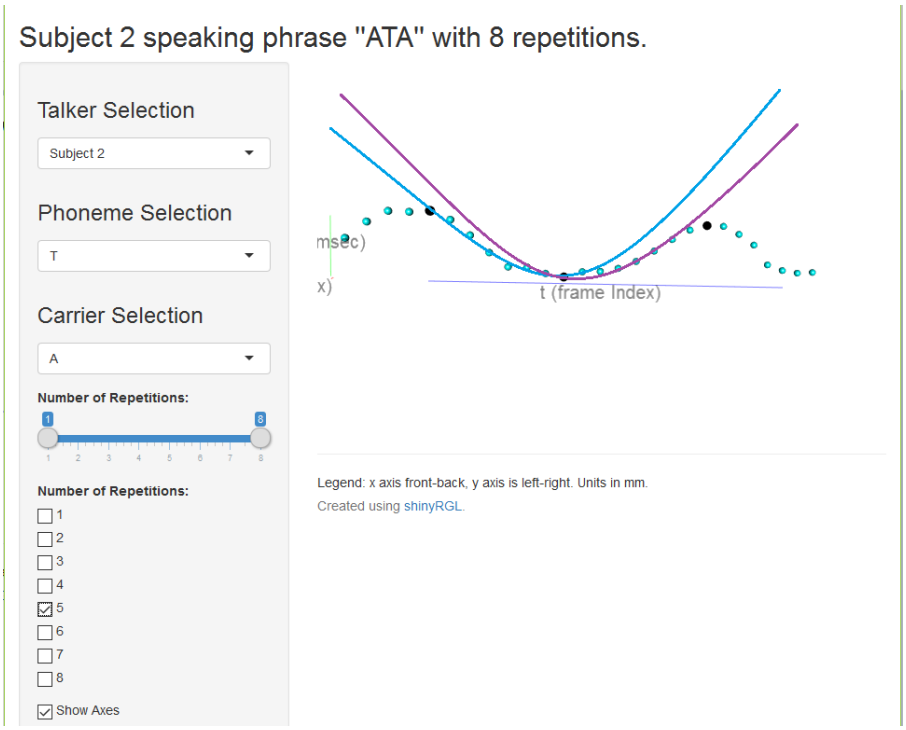

Figure 3.6: Fitting two polynomial of degree 2 to the deceleration segment (fitted polynomial shown in blue) and the acceleration segment (fitted polynomial shown in magenta)

each observed trajectory. A screenshot is shown in Figure 3.6.

3.5.3 Feature Normalization

Once the features were derived for all of the speed profiles, they were normalized to [0,1]. Normalization was done using the following procedure. First, the minimum and maximum value from all a_i coefficients of all segments was identified, min(a_i) and max(a_i) respectively. Then the following formula was used:

normalized coeficient value C_i = (C_i-min(C))/(max(C)-min(C)) where C=(C_1,...,C_n) and n=

the number of segments

3.5.4 Feature Set 2

Feature set 2 is also a set of 6 features that was extracted from each speed profile, albeit using a slightly modified technique. Like feature set 1, features set 2 also depends on coefficients of two polynomial curves that were fit to the speed profile. However, the polynomials were fitted to a set of points that was constructed slightly differently to consists of three points for each segment. The deceleration and acceleration phases were identified, as in the previous case.

In the case of the first (deceleration) segment, the first two relevant points are the first maximum point and the minimum point. Then a third "reflected point" was added, to mirror the first maximum point, using a vertical axis of symmetry that was positioned at the minimum point.

For the second, acceleration segment, the second and third points are the minimum point and the second maximum point. For the first point, a " reflected point" was added to mirror the second maximum point, using a vertical axis of symmetry that was positioned at the minimum point. This "reflected point" for the acceleration is illustrated in Figure 3.7.

Two polynomials of degree 2, $y = a_1x^2 + b_1x + c_1$ and $y = a_2x^2 + b_2x + c_2$ were fit to each of these three-point sets and the coefficients of the two parabola were derived, $(a1, b1, c1)$ and $(a2, b2, c2)$, respectively. The result is a total of 6 features for each speed profile.

Once the features were derived for all of the speed profiles, they were normalized. The procedure of normalization was described in section 3.5.3.

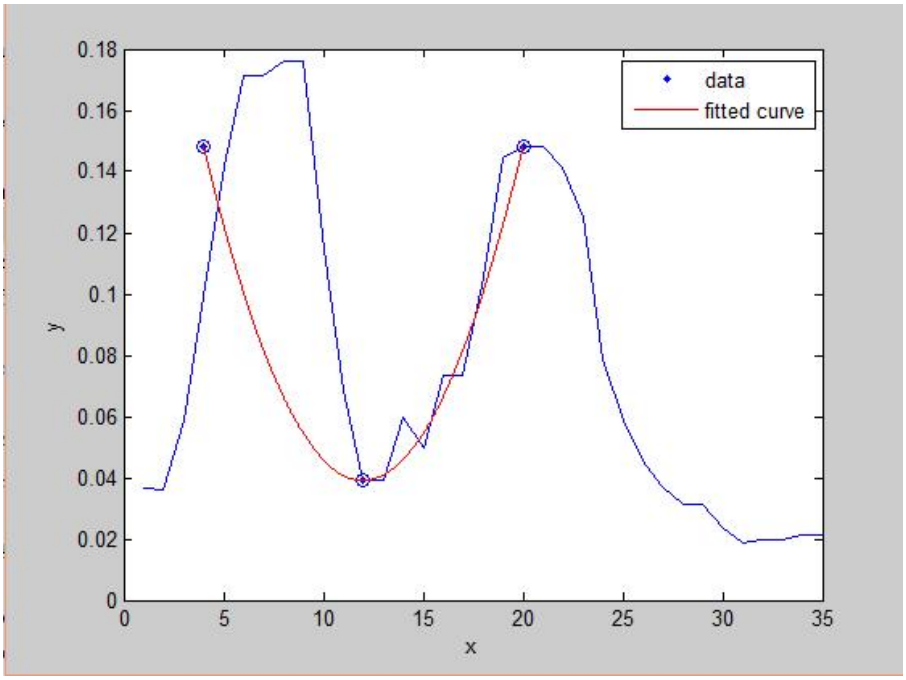

Figure 3.7: A polynomial (red curve) fitted to the acceleration segment, using a set of three points (in order from left to right, the reflection of second maximum point, the minimum point and the second maximum point).

Once the features were derived for all of the speed profiles, they were normalized. The procedure of normalization is described in previous section.

3.6 Classification Approaches

In the next section, I will describe a series of machine learning experiments. First, I will briefly review the two machine learning techniques that I will employ, both of which were summarized in detail in section 5 of chapter 2.

3.6.1 SVM classification

To perform SVM classification, I used the implementation that is provided in Matlab (Canu et al. 2005). The feature sets as described in the preceding section and the consonant segment labels were considered as the input of SVM classifier. Multi-class classification using SVM was done using the one-against-one strategy. Sequential Minimal Optimization (SMO), which is the default method for optimization in the SVM classification, must contain exactly two groups to classify, but there are up to eight classes considered in this work, therefore I used Least Squares (LS) as the optimization method to find the separating hyperplane. I used the linear kernel for the decision function.

3.6.2 HMM classification

To perform HMM classification, I used the HMM package within the Waikato Environment for Knowledge Analysis (WEKA) software suite (Hall et al. 2009). I determined the numbers of

states and folds empirically, through experimenting with a number of different configurations.

Once the parameters were determined, I trained eight different HMMs, one HMM per class. To perform classification on a feature set series with an unknown label, I ran a forward algorithm for each HMM and received the probability value P(HMM | observation) for each one. The classification result was taken on the basis of the result with the highest probability.

3.7 Studies

3.7.1 Tasks

In the studies below, I make use of the following four types of discrimination tasks:

1. Task 1: distinguish between one plosive and one fricative VCVs (/t/, /s/)
2. Task 2: distinguish among plosives (/t/, /g/, /k/, /d/)
3. Task 3: distinguish among fricatives (/s/, /z/, /sh/) and one affricate (/tch/)
4. Task 4: distinguish among all 8 consonant segment classes (/s/, /tch/, /z/, /sh/, /t/, /g/, /k/, /d/).

For task 1, all pairwise combinations of plosive and fricative VCVs are shown in table 3.2.

All Plosive and Fricative Pairings
ATA-ASA
ATA-AZA
ATA-ASHA
ADA-ASA
ADA-AZA
ADA-ASHA
AGA-ASA
AGA-AZA
AGA-ASHA

Table 3.2: All possible plosive and fricative pairings, for Task 1

For task 2, all pairwise combinations of plosives VCVs are shown in table 3.3.

All pairwise combinations of plosives (Task 2)
ATA-ADA
ATA-AKA
ATA-AGA
ADA-AKA
ADA-AGA
AKA-AGA

Table 3.3: All possible pairwise combinations of plosives, for task 2.

For task 3, all pairwise combinations of fricatives VCVs are shown in table 3.4.

All pairwise combinations of fricatives and affricates (Task 3)
ASA-AZA
ASA-ASHA
AZA-ASHA

Table 3.4: All possible pairwise combinations of fricatives and affricates for task 3

The table of all pairwise combinations of the full set of 8 consonant segments (/s/, /tch/, /z/, /sh/, /t/, /g/, /k/,/d/) is omitted for brevity.

3.7.2 Study 1 Design and Dataset Preparation

The objective of study 1 is to determine the following: Given a talker ‘s own data concerning tongue-tip kinematic profiles for different consonant segments, how accurately can we classify the consonant segment of unknown tongue-tip kinematic profile by that same talker?

To answer this question, I posed four different types of discrimination tasks:

1. Task 1: distinguish between one plosive and one fricative VCVs (/t/, /s/)
2. Task 2: distinguish among plosives (/t/, /g/, /k/, /d/)
3. Task 3: distinguish among fricatives (/s/, /z/, /sh/) and one affricate (/tch/)
4. Task 4: distinguish among all 8 consonant segment classes

For each task, we need to derive classifier accuracy for each of 17 different sub-experiments (one for each speaker). I plan to employ each of SVM and HMM for the classifier techniques.

Study 1 consists of per-talker task evaluation. Each task requires its own set of 17 sub-experiments, one per talker. Each of these 17 sub-experiments requires its own talker-specific sub-datasets: one sub-dataset for the SVM classifier and another sub-dataset for the HMM classifier. The SVM classifier will employ the speed-based features (as described in section 1.4) and the HMM classification will employ the velocity-based features (as described in section 3.4).

For SVM classification, 4-fold cross-validation will be used, meaning that the data set is divided into 4 subsets, and the classification is repeated 4 times. For each fold, training is based on a three-subsets of the sub-dataset (75%) and testing on the remaining subset of the dataset (25%). 4-fold cross validation was used to ensure that all of the data has been used in both training and testing data sets. Then the average across all 4 trials is computed.

For each task described below, one experiment will be performed for each talker.

1. Task 1, Distinguishing between Plosives and Fricatives: 17 sub-datasets were prepared constructed on a per-talker basis. Each talker-specific sub-dataset consists of speed features from feature sets 1 and 2 (for SVM) and of the velocity features (for HMM) of a particular talker's 8 repetitions of one plosive VCV and 8 repetitions of one fricative VCV. Each feature vector is labeled as one plosive or one fricative.

2. Task 2, Distinguishing among Plosives: 17 talker-specific sub-datasets were constructed. For the SVM classifier, each sub-dataset consists of features from feature sets 1 and 2 of a particular talker‘s 8 repetitions of all the plosive VCVs, each labeled according to type: /t/, /g/, /k/, /d/.

   For the HMM classifier, each sub-dataset consists of the velocity sequences of a particular talker‘s 8 repetitions of all plosive VCVs.

3. Task 3, Distinguishing among fricatives: 17 talker-specific sub-datasets were constructed. For the SVM classifier, each sub-dataset consists of features from feature sets 1 and 2

of a particular talker‘s 8 repetitions of all the fricative and the one affricate VCVs, each labeled according to type: /s/, /z/, /sh/ and /tch/. Affricates are pooled with fricatives in this experiment.

For the HMM classifier, each sub-dataset for HMM consists of the velocity sequences of a particular talker‘s 8 repetitions of all the fricative and the affricate VCVs.

4. Task 4, Distinguishing among all consonant segments: 17 talker-specific sub-datasets were constructed. For the SVM classifier, each sub-dataset consists of the speed-based features from feature sets 1 and 2, and, for the HMM classifier, each sub-dataset consists of the velocity sequences, of a particular talker ’s 8 repetitions of all 8 types of consonant segments, each labeled according to type: (/s/, /tch/, /z/, /sh/, /t/, /g/, /k/, /d/).

3.7.3 Study 2 Design and Dataset Preparation

The objective of study 2 is to determine, given data representing the tongue-tip kinematic profiles for a pool of different talkers, how accurately can we classify the consonant segment of unknown tongue-tip kinematic profile by a talker from outside that pool?

To answer this question, the same set of four discrimination tasks was posed as in Study 1:

1. Task 1: distinguish between plosive and fricative VCVs (/t/, /s/)
2. Task 2: distinguish among the plosives (/t/, /g/, /k/, /d/)
3. Task 3: distinguish among the fricatives (/s/, /z/, /sh/) and the affricate (/tch/)
4. Task 4: distinguish among all 8 consonant segment classes

For each task, we need to derive classifier accuracy for each of sub-experiments (one for each SVM and HMM classifier). I plan to employ each of SVM and HMM for the classifier techniques.

Study 2 consists of task evaluation using a leave-one-out approach. Each task requires its own set of 17 sub-experiments, one sub-experiment per left-out talker. Each of these 17 sub-experiments requires its own talker-specific sub-datasets: one sub-dataset for the SVM classifier and another sub-dataset for the HMM classifier. These sub-datasets are constructed on a per-task basis as described in Study 1, except that the training data for each sub-experiment consists of the pooled data of all the talkers, except the left-out talker (16 talkers). The left-out talker is used for testing.

3.7.4 Study 3 Design and Dataset Preparation

The objective of study 3 is to determine, given data representing the tongue-tip speed profiles for a pool of different talkers: how accurately can we classify the consonant segment of unknown tongue-tip trajectories by a talker from within that pool?

This study design is very similar to the study 1 and study 2. The same tasks were posed. Whereas study 1 and 2 required 17 sub-experiments for each task (one sub-experiment for each talker, either using per-talker or leave-one-out training), study 3 requires one data set per task. The same type of datasets were used for each task (for SVM and HMM respectively, except all talker data was pooled together and training/testing splits used. Folds were drawn over all speakers and I used four-fold cross-validation. The training set was composed of 75% of the data which reflects 75% of each speaker's repetitions. Testing was performed using the other 25% of

the repetitions.

For the HMM experiments, I used a variety of different folds. By choosing 10 fold cross validation, Weka takes 100 labeled data and produces 10 equally sized sets, then divided into two groups: 90 labeled data are used for training and 10 labeled data are used for testing.

3.8 Conclusion

In this chapter, I outlined the three questions of interest in this work. I described the suite of three studies that I designed in order to answer these questions, and the data collection and preparation procedure for each study. The study design entails the use of both SVM and HMM classification, and each study makes use of datasets that employ different sets of feature, which have been derived from the speed and velocity properties of the kinematic data.

Chapter 4

Study Results

In this chapter I will present the results of the studies that were performed in order to answer following three questions:

1. Question 1: Given a talker's own data concerning tongue-tip speed profiles for different consonant segments, how accurately can we classify the consonant segment type of unknown tongue-tip speed profile by that same talker?

2. Question 2: Given data representing the tongue-tip speed profiles for a pool of different talkers, how accurately can we classify the consonant segment type of unknown tongue-tip trajectories by a talker from outside that pool?

3. Question 3: Given data representing the tongue-tip speed profiles for a pool of different talkers, how accurately can we classify the consonant segment type of unknown tongue-tip trajectories by a talker from within that pool?

4.1 Study 1 Results

Question 1: Given a talker 's own data concerning tongue-tip speed profiles for different consonant segments, how accurately can we classify the consonant segment of unknown tongue-tip speed profile by that same talker?

This study entailed the following four tasks:

1. Task 1: distinguish between plosives and fricatives
2. Task 2: distinguish among plosives (/t/, /d/, /k/, /g/)
3. Task 3: distinguish among fricatives (/s/, /sh/, /z/)
4. Task 4: distinguish among all 8 consonant segments (over three classes: plosives, fricatives, affricates)

4.1.1 Study 1, Task 1: Distinguishing between Plosives and Fricatives

This task was performed using each of two different feature sets and the SVM classification technique. The HMM classification technique was employed using the third feature set (the three different feature sets were described in the previous chapter). In all the following experiments, training was based solely on a talker‘s own data.

4.1.1.1 Study 1, Task 1: SVM Classification on the basis of Feature Set 1

Speaker-specific results were derived by considering each of the possible pairwise combinations between the plosives (/t/, /d/, /g/) and the fricatives (/s/, /z/, /sh/). The plosive /k/ was omitted.

Feature set 1 is a set of six features which was extracted from each speed profile. These features are drawn directly from the three coefficients of two polynomial curves that were fit to the deceleration and acceleration segments of each speed profile. Each consonant was classified in the context of a vowel-consonant-vowel (VCV) segment (the "A*A" VCV segment, specifically). The classification accuracies per pairwise combination and the mean accuracy that was derived over the 9 possible pairings are presented in table 4.1.

SVM classification accuracy between Plosive-fricative VCVs using feature 1										
Speaker/P-F	ATA-ASA	ATA-AZA	ATA-ASHA	ADA-ASA	ADA-AZA	ADA-ASHA	AGA-ASA	AGA-AZA	AGA-ASHA	Mean
W02	100	95.75	–	100	96.5	–	100	92.5	–	97.45
W07	93	100	93.25	100	96.75	93.75	93.75	100	93.5	98
W17	100	93.75	100	100	100	100	93.75	96.75	93.75	97.55
W24	100	92.75	93.75	94.22	100	93.75	100	100	96.75	92
W21	100	100	100	93.75	93.75	91.5	91.25	100	91.5	95.75
W09	90.75	100	86.25	91.75	100	81.25	100	100	83.75	92.63
W22	91.5	96.25	100	100	100	100	79	96.75	96.75	95.58
W13	97.5	87.5	91.5	100	93.75	100	87.5	66.75	100	91.61
W14	66.5	81.25	81.25	100	93.5	91.5	93.75	93.75	100	89.05
W11	92.5	93.75	86.25	100	87.5	93.75	94.75	72.25	72.75	84
W20	100	63.25	90.25	100	100	100	91.75	86	79.75	94
W10	83.25	72.3	100	93.75	85.5	100	87.25	93.75	93.75	89.95
W25	92.75	93.75	100	100	93.75	100	75.25	72.25	89.75	90.83
W23	93.75	100	81	93.75	100	93.75	93.5	98.75	74.75	92.13
W15	100	65	100	–	–	–	91.5	72.75	100	88.20
W12	91.5	70.5	82	60.25	79	100	69	73	75	77.80
W19	90.5	71.75	93.75	53.75	93.75	93.75	72.75	68	70.75	78.75
Mean	93.14	86.91	92.45	92.57	94.60	95.53	89.10	87.25	88.28	90.90

Table 4.1: SVM classification accuracy between Plosive-fricative VCVs using feature set 1

The mean classification accuracy between plosive-fricative VCVs over all the talkers was 90.90%. Talkers W07, with 98%, and W12, with 77.80%, had the best and worst mean accuracies

for consonant segment classification, respectively. The most accurately distinguished pair was ADA-ASHA (95.53%) and the least accurately distinguished pair was ATA-AZA (86.91%).

4.1.1.2 Study 1, Task 1: SVM Classification on the basis of Feature Set 2

The same experiment was performed with feature set 2. Feature set 2 is also a set of 6 features that was extracted from each speed profile, albeit using a slightly modified technique (see section 3.1.1.1). The classification accuracies are presented in table 4.2.

SVM classification accuracy between Plosive-fricative VCVs using feature set 2										
Speaker/P-F	ATA-ASA	ATA-AZA	ATA-ASHA	ADA-ASA	ADA-AZA	ADA-ASHA	AGA-ASA	AGA-AZA	AGA-ASHA	Mean
W02	100	93.75	–	100	93.5	–	100	87.5	–	95.79
W07	100	100	93.75	100	93.75	93.75	93.75	100	87.5	94
W17	100	93.75	100	100	100	100	93.75	93.75	93.75	97.22
W24	93.75	93.75	93.75	100	100	93.75	100	100	93.75	96.52
W21	100	100	100	93.75	93.75	87.5	87.5	100	87.5	94.44
W09	93.75	100	81.25	93.75	100	81.25	100	100	77.25	91.91
W22	87.5	93.75	100	100	100	100	75	93.75	93.75	90
W13	100	87.5	91.5	100	93.75	100	87.5	64.75	100	91.66
W14	62.5	81.25	79	100	91.5	91.5	93.75	93.75	100	85
W11	100	93.75	81.25	100	87.5	93.75	93.75	72.25	68.75	87.88
W20	100	59.5	85.25	100	100	100	91.5	83	76.75	88.44
W10	81.25	62.5	100	93.75	87.5	100	83.75	93.75	93.75	95
W25	87.25	93.75	100	100	93.75	100	71.25	68.25	85.25	88.83
W23	93.75	100	75	93.75	100	93.75	87.5	93.75	68.75	89.58
W15	100	60	100	–	–	–	91.5	68.75	100	86.70
W12	87.5	70.5	75	60.25	79	100	63	63	75	74.80
W19	87.5	68.75	93.75	53.75	93.75	93.75	68.75	64	68.75	76.97
Mean	92.632	85.44	90.59	93.06	94.23	95.26	87.19	84.72	85.65	89.69

Table 4.2: SVM classification accuracy between Plosive-fricative VCVs using feature set 2

The mean classification accuracy was 89.69 % over all the talkers. Once again, W17 with

97.22% and W12 with 74.80% have the best and worst mean accuracies, respectively. The most accurately distinguished pair was ADA-ASHA with 95.26% accuracy. The least accurately distinguished pair was ATA-AZA with 85.44%.

Talker W12, whether using either feature set 1 or feature set 2, had the worst classification accuracy. The plosive-fricative pair ADA-ASHA was the most distinguishable, with 95.53% and 92.26%, respectively, for each of feature set 1 and feature set 2.

4.1.1.3 Study 1, Task 1: HMM Classification on the basis of Feature Set 3

HMM classification accuracies between Plosive-Fricative VCVs are presented in table 4.3:

The mean classification accuracies between plosive-fricative VCVs over all the talkers was 93.51%. Talkers W02, with 99.20%, and W19, with 83.10%, have the best and worst mean accuracies for consonant segment classification, respectively. The most accurately distinguished pair was ADA-ASHA (97.71%), and the least accurately distinguished pair was AGA-AZA (89.40%).

4.1.1.4 Study 1, Task 1 Summary

I next compare the Study 1, Task 1 results to examine the impact of classification technique and feature set. The per-talker data is displayed in figure 4.1.

The mean accuracy over all the talkers and all pairing using the HMM technique was 93.10%, which was greater than SVM1 (90.75%) and SVM2 (89.97%). The difference in SVM classification accuracies using feature set 1 and the accuracies using feature set 2 is not significant level ($t(17) = -0.57123$, $p = 0.285916$).

HMM classification accuracy between Plosive-fricative VCVs										
Speaker/P-F	ATA-ASA	ATA-AZA	ATA-ASHA	ADA-ASA	ADA-AZA	ADA-ASHA	AGA-ASA	AGA-AZA	AGA-ASHA	Mean
W02	100	100	–	100	100	–	100	95.25	–	99.20
W07	100	100	100	100	93.75	100	93.75	100	95.5	98.11
W17	100	93.75	100	100	100	100	93.75	93.75	100	97.91
W24	93.75	100	93.75	100	100	93.75	100	100	98.75	97.77
W21	100	100	100	93.75	100	90.5	93.5	100	95.5	97.02
W09	100	100	93.25	93.75	100	90.25	100	100	85.25	95.83
W22	94.5	93.75	100	100	100	100	83	93.75	93.75	95.41
W13	100	93.5	96.5	100	100	100	87.5	75.25	100	94.75
W14	78.5	89.75	86.25	100	96.5	97.5	100	93.75	100	93.58
W11	100	98.75	90.25	100	95.5	93.75	100	78.25	80.75	93.02
W20	100	71.5	93.25	100	100	100	91.5	92	84.75	92.55
W10	90.25	80.5	100	93.75	87.5	100	83.75	100	93.75	92.16
W25	92.75	93.75	100	100	93.75	100	79.25	75.23	91.25	91.77
W23	96.75	100	82	93.75	100	100	87.5	93.75	68.75	91.38
W15	100	73.5	100	–	–	–	91.5	81.75	100	91.12
W12	87.5	77.5	75	72.25	87	100	76.33	75	83.25	85
W19	93.25	75.75	98.75	65.75	96.70	100	76.75	72.23	68.75	83.10
Mean	95.72	90.70	94.31	94.56	96.91	97.71	90.47	89.40	90	93.51

Table 4.3: HMM classification accuracy between Plosive-fricative VCVs

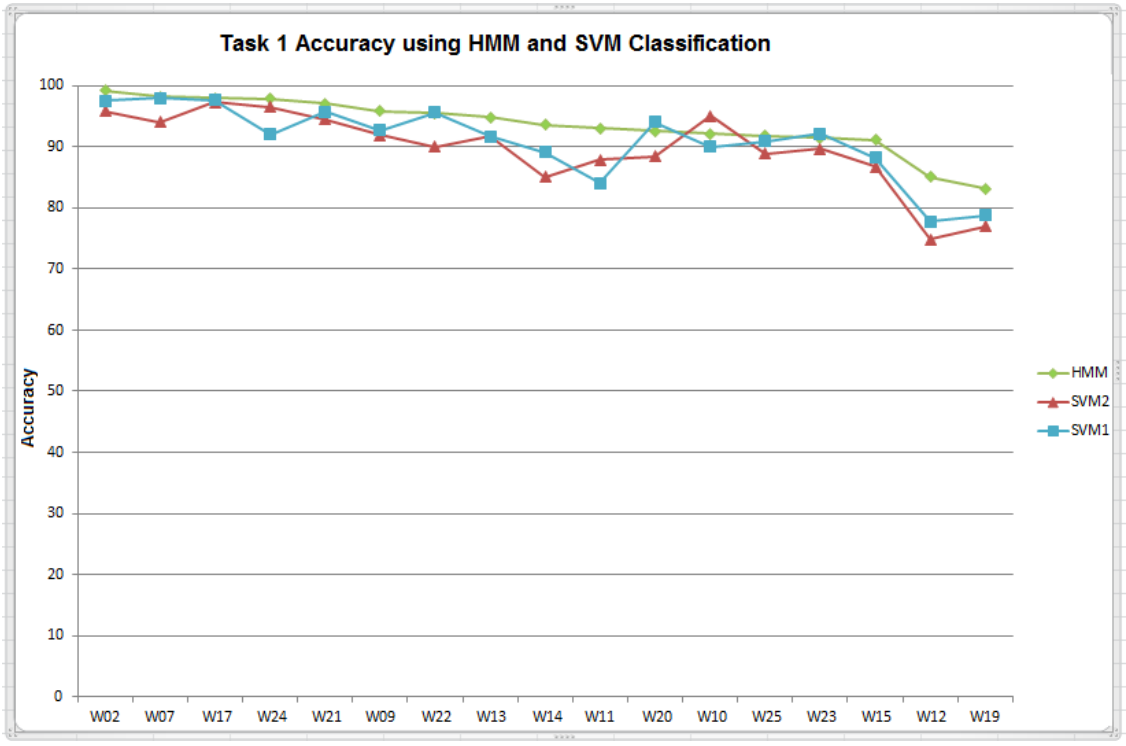

Figure 4.1: Accuracy results for plosive vs fricative classification for all three classification tasks. Talkers are arranged in order of descending accuracy based on HMM results.

4.1.2 Study 1, Task 2: Distinguishing among Plosives

This task was performed using each of two different feature sets and the SVM classification technique. The HMM classification technique employed the third feature set.

The results were derived by considering each of the six possible pairwise combinations among the plosives: /t/, /d/, /g/. The mean accuracy was derived over these 6 possible pairings. In these experiments, training was based solely on a talker‘s own data.

4.1.2.1 Study 1, Task 2: SVM classification on the basis of Feature Set 1

The classification accuracies obtained using the SVM classification technique using feature set 1 are presented in table 4.4.

The mean classification accuracies among plosive VCVs over all the talkers was 85.94%. Talkers W13, with 100%, and W11, with 75%, have the best and worst mean accuracies for consonant segment classification, respectively. The most accurately distinguished pair was ATA-ADA (89.69%) and the least accurately distinguished pair was AKA-AGA (74.45%).

4.1.2.2 Study 1, Task 2: SVM classification on the basis of Feature Set 2

The classification accuracies obtained using the SVM technique using feature set 2 are presented in table 4.5.

The mean classification accuracies among plosive VCVs over all the talkers was 81.97%. Talkers W13, with 100%, and W25, with 75.35%, have the best and worst mean accuracies for consonant segment classification, respectively. The most accurately distinguished pair was ADA-AKA

SVM classification accuracy among plosives using feature set 1							
Speaker/P-P	ATA-ADA	ATA-AKA	ATA-AGA	ADA-AKA	ADA-AGA	AKA-AGA	Mean
W13	100	–	100	–	100	–	100
W22	87.66	85.25	89.25	96.75	100	89.25	91.36
W21	93.5	100	91.5	100	89.25	79	92.20
W07	95.75	93.75	86.25	100	80	100	92.62
W24	91.5	80.75	93.75	100	93.75	75	89.12
W02	92.5	85.25	92.5	95.25	72.33	96.75	89.09
W15	–	92.5	94.5	–	–	72	82
W20	80.25	76.75	87	89.33	100	86	85
W14	96.5	85.25	87.25	96.5	71.5	73	80
W23	79	100	64	100	85.25	86.25	92.3
W09	81.25	84.25	91.5	79	71.5	84	81.91
W19	93.75	100	79	91.5	91.5	52	84.62
W10	79	92.5	96.75	75	81.25	68.75	80
W17	93.5	92.25	79	75	93.75	51	80.75
W25	91.25	89.25	69	92.25	80	64.26	89
W12	92.5	66.5	57	66	75.25	52	76
W11	87.25	62	62.5	75	59.25	62	75
Mean	89.69	86.64	83.57	88.77	84.03	74.45	85.94

Table 4.4: SVM classification accuracy amongn plosives using feature set 1.

SVM classification accuracy among Plosives using feature set 2							
Speaker/P-P	ATA-ADA	ATA-AKA	ATA-AGA	ADA-AKA	ADA-AGA	AKA-AGA	Mean
W13	100	–	100	–	100	–	100
W22	81.25	81.25	81.25	93.75	93.75	81.25	85.41
W21	91.5	100	91.5	93.75	87.5	73	89.54
W07	93.75	93.75	81.25	100	75	100	90.625
W24	87.5	76.75	93.75	100	93.75	71	87.12
W02	87.5	81.25	87.5	93.75	66.5	93.75	85.04
W15	–	87.5	91.5	–	–	69	82.66
W20	81.25	76.75	83	85.25	100	83	84.87
W14	91.5	87.5	81.25	91.5	66.5	69	81.20
W23	75	100	58	100	81.25	81.25	82.58
W09	81.25	81.25	87.5	79	66.5	80	79.25
W19	93.75	100	75	87.5	87.5	45	81.45
W10	75	87.5	93.75	75	81.25	68.75	80.20
W17	87.5	87.5	79	79	93.75	42	78.12
W25	87.75	85.25	59	85.25	80	55	75.37
W12	87.5	62.5	53	63	68.75	57	65.29
W11	81.25	58	62.5	75	53	59	64.79
Mean	86.45	84.17	79.92	86.78	80.93	70.5	81.97

Table 4.5: SVM classification accuracy among plosives using feature set 2

(86.78%) and the least accurately distinguished pair was AKA-AGA (70.5%).

Talker W13, using either feature set 1 or feature set 2, had the best classification accuracy. The plosive pair AKA-AGA was the least distinguishable, with 74.45% and 70.5% accuracy, respectively, for each of feature set 1 and feature set 2.

4.1.2.3 Study 1, Task 2: HMM classification on the basis of Feature Set 3

HMM classification accuracies among plosive VCVs are presented in table 4.6.

HMM classification accuracy among Plosives							
Speaker/P-P	ATA-ADA	ATA-AKA	ATA-AGA	ADA-AKA	ADA-AGA	AKA-AGA	Mean
W13	100	–	100	–	100	–	96.79
W22	94.26	90.25	94.25	100	100	96.66	95.90
W21	100	100	95.5	100	95.75	86.33	96.26
W07	100	96.75	90.25	100	85	100	95.33
W24	96.5	89.75	98.75	100	100	83	94.66
W02	96.5	89.75	96.5	98.33	80.33	97.75	93.193
W15	–	96.5	98.75	–	–	81	92.083
W20	86.25	81.75	93.25	95.33	100	94	91.76
W14	100	90.25	93.25	100	82.5	79.33	90.88
W23	85	100	70.23	100	91.25	93.75	90.03
W09	86.25	89.75	96.5	85	78.25	88	87.29
W19	96.75	100	85	95.5	94.25	60.33	88.63
W10	83	96.5	100	82.33	89.25	73.75	87.47
W17	96.5	100	85	81	93.75	62	86.375
W25	96.65	93.25	75	92.25	87.42	72.26	86.13
W12	96.5	71.5	65	72.33	84.25	60	87.2
W11	93.25	68	71.5	82.25	65.25	71.33	87
Mean	94.21	90.875	88.74	92.28	89.20	81.21	90.95

Table 4.6: HMM classification accuracy among plosives.

The mean classification accuracies among plosive VCVs over all the talkers was 90.95%. Talk-

ers W13, with 96.76%, and W25, with 86.13%, have the best and worst mean accuracies for consonant segment classification, respectively. The most accurately distinguished pair was ATA-ADA (94.21%). The least accurately distinguished pair was AKA-AGA (81.21%).

4.1.2.4 Study 1, Task 2 Summary

I next compare the Study 1, Task 1 results to examine the impact of classification technique and feature set. The per-talker data is displayed in figure 4.2.

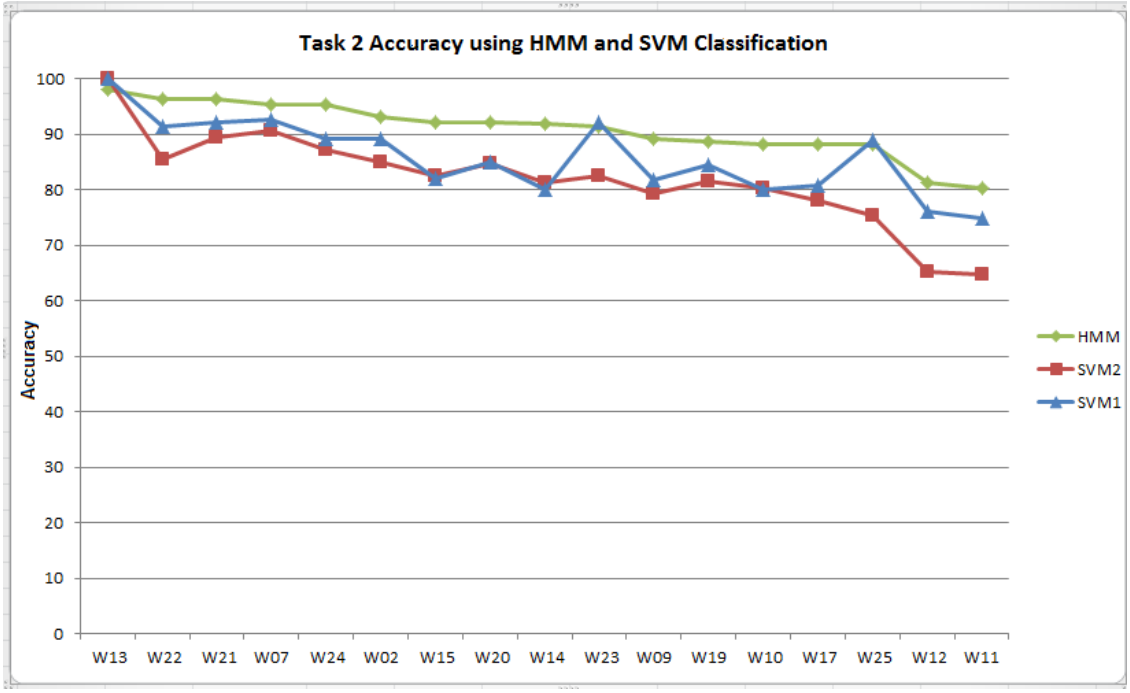

Figure 4.2: Accuracy results for plosives classification for all three classification tasks. Talkers are arranged in order of descending accuracy based on HMM results.

The mean accuracy over all the talkers and all pairing using the HMM technique was 90.95%, which is greater than SVM1 (85.94%) and SVM2 (81.97%). SVM classification accuracy using

feature set 1 and feature set 2 was not significant at the $p<0.05$ level ($t(n) = 1.50511$, $p = 0.071052$).

4.1.3 Study 1, Task 3: Distinguishing among Fricatives

This task was performed using each of two different feature sets and the SVM classification technique. The HMM classification technique used the third feature set.

The results were derived by considering each of the six possible pairwise combinations among the fricatives: /s/, /z/, /sh/. The mean accuracy was derived over these 6 possible pairings. In these experiments, training was based solely on a talker's own data.

4.1.3.1 Study 1, Task 3: SVM classification on the basis of Feature Set 1

The classification accuracies obtained using the SVM technique using feature set 1 to distinguish among fricatives are presented in table 4.7.

The mean classification accuracies among fricative VCVs over all the talkers was 77.93%. Talkers W10, with 96.83%, and W20, with 69.41%, have the best and worst mean accuracies for consonant segment classification, respectively. The most accurately distinguished pair was AZA-ASHA (80.27%). The least accurately distinguished pair was ASA-ASHA (75.70%).

4.1.3.2 Study 1, Task 3: SVM classification using Feature Set 2

The classification accuracies obtained using feature set 2 are presented in table 4.8.

Distinguishing between fricatives AZA and ASHA, on the basis of either feature set 1 or feature set 2, had the best accuracy over all the possible fricative pairings, with 80.27% and

SVM classification accuracy among fricatives VCVs using feature set 1				
Speaker/F-F	ASA-AZA	ASA-ASHA	AZA-ASHA	Mean
W17	89	83	65.6	79.2
W11	100	93.75	78.25	90.66
W10	96.75	93.75	100	96.83
W24	87.25	65.75	65.25	72.75
W22	93.75	48	100	80.58
W25	93.76	88.33	95.75	92.613
W14	88.5	53.2	86	75.9
W23	100	76.5	88	88.16
W12	75.25	68.25	68.75	70.75
W15	64	75.5	78.75	72.75
W21	81	88	100	89.66
W13	73	88.75	70.33	77.36
W20	62	88	58.25	69.41
W07	52	70	78.25	66.75
W19	64.33	78.25	78	73.52
W09	61.75	52.25	73.25	62.41
W02	65.5	–	–	–
Mean*	79.28	75.705	80.27	77.93

* Mean over all 17 speakers. For W02, single value was used in place of the mean value.

Table 4.7: SVM classification accuracy among fricatives VCVs using feature set 1

SVM classification accuracy among fricatives VCVs using feature set 2				
Speaker/F-F	ASA-AZA	ASA-ASHA	AZA-ASHA	Mean
W17	93.75	93.75	100	95.83
W11	100	93.75	81.25	91.66
W10	93.75	87.5	100	93.75
W24	81.25	81.25	81.25	81.25
W22	93.75	81.25	87.5	87.5
W25	88.76	81.25	93.75	87.92
W14	87.5	71	100	86.16
W23	100	63	93.75	85.58
W12	85.25	73	66	74.75
W15	59	93.75	93.75	82.16
W21	75	63	81.25	73.08
W13	70	75	73	72.66
W20	55	80	91.75	75.58
W07	55	75	72.75	67.58
W19	63	69	63	65
W09	60.25	50	81.25	63.83
W02	66.5	–	–	–
Mean*	78.10	76.96	85.01	79.46

* Mean over all 17 speakers. For W02, single value was used in place of the mean value.

Table 4.8: SVM classification accuracy among fricatives VCVs using feature set 2

85.27% accuracy, respectively.

The mean classification accuracies among fricative VCVs over all the talkers was 79.46%. Talkers W17, with 95.83%, and W19, with 65%, had the best and worst mean accuracies for consonant segment classification, respectively. The most accurately distinguished pair was AZA-ASHA (85.27%). The least accurately distinguished pair was ASA-ASHA (76.96%).

The fricative pairs AZA-ASHA was the best distinguished, with 74.45% and 80.27% accuracy, for each of feature set 1 and feature set 2, respectively. The fricative pairs, ASA-ASHA was the least accurately distinguished pair, with 75.70% and 76.96% for each of feature set 1 and feature set 2, respectively.

4.1.3.3 Study 1, Task 3: HMM Classification Using Feature Set 3

HMM classification accuracies among fricative VCVs using feature set 3 are presented in table 4.9.

The mean classification accuracies among fricative VCVs over all the talkers was 90.92%. Talkers W17, with 100%, and W02, with 78%, have the best and worst mean accuracies for consonant segment classification, respectively. The most accurately distinguished pair was AZA-ASHA (94.38%) and the least accurately distinguished pair was ASA-AZA (88.48%).

HMM classification accuracy among fricatives				
Speaker/F-F	ASA-AZA	ASA-ASHA	AZA-ASHA	Mean
W17	100	100	100	100
W11	95.6	100	85.25	99
W10	100	96.25	100	98.75
W24	94.25	98.75	97.5	96.83
W22	100	90.25	100	96.75
W25	93.76	92.5	100	95.42
W14	100	84.75	100	94.91
W23	100	80.33	100	93.44
W12	97.25	85.25	79.75	93
W15	73.33	100	100	91.11
W21	89	78	95.5	89.5
W13	86	87.25	92.33	88.52
W20	71.25	93.75	100	88.33
W07	72.25	90.25	86.25	82.91
W19	80.33	83.5	79	80.94
W09	73.25	67	94.5	78.25
W02	78	–	–	–
Mean*	88.48	89.23	94.38	90.92

* Mean over all 17 speakers. For W02, single value was used in place of the mean value.

Table 4.9: HMM classification accuracy among fricatives

4.1.3.4 Study 1, Task 3 Summary

I next compare the Study 1, Task 3 results to examine the impact of classification technique and feature set. The per-talker data is displayed in figure 4.3.

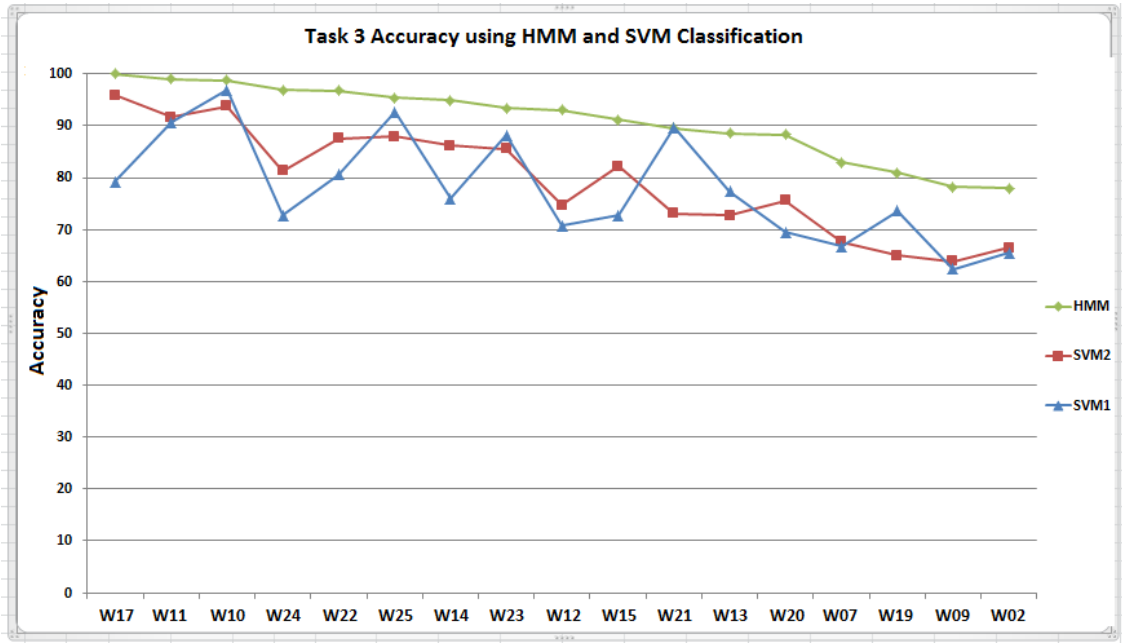

Figure 4.3: Accuracy results for fricatives classification for all three classification tasks. Talkers are arranged in order of descending accuracy based on HMM results.

The mean accuracy over all the talkers and all pairing using the HMM technique was 90.92%, which was greater than SVM1 (77.93%) and SVM2 (79.46%). The SVM classification accuracy using feature set 1 and feature set 2 was not significant at the $p<0.05$ level ($t(n) = -0.42954$, $p = 0.335206$).

4.1.4 Study 1, Task 4: Distinguishing among all 8 Consonants

4.1.4.1 Study 1, Task 4: All Classification Results

The results derived using each of the SVM1, SVM2 and HMM classification approaches (each employing their respective feature sets of set 1, set 2 and set 3) are provided in table 4.10.

SVM1, SVM2 and HMM classification accuracy among all 8 classes			
Speaker/approach	SVM1	SVM2	HMM
W24	63.5	75.5	91.4
W02	60.3	65.81	90.5
W17	67.25	68	90.5
W21	65.5	60.3	88.5
W10	52	60.3	87.7
W23	65.5	66.3	86.7
W13	55	61.5	86.3
W15	69	62	85.75
W11	65	61	85.2
W07	60.2	58.5	83.5
W14	70.75	61.25	81.6
W12	65.5	55.6	80
W09	59.25	56.5	79.3
W20	67.5	52.5	79.2
W25	60.25	56.33	77.65
W22	59.65	53.2	76.3
W19	69	55.25	75
Mean	63.24	60.57	82.32

Table 4.10: SVM1, SVM2 and HMM classification accuracies distinguishing among the 8 consonant classes

The mean accuracies of HMM, SVM1 and SVM2 were 82.3%, 63.2% and 60.57% respectively.

I next compare the Study 1, Task 4 results to examine the impact of classification technique and feature set. The per-talker data is displayed in figure 4.4.

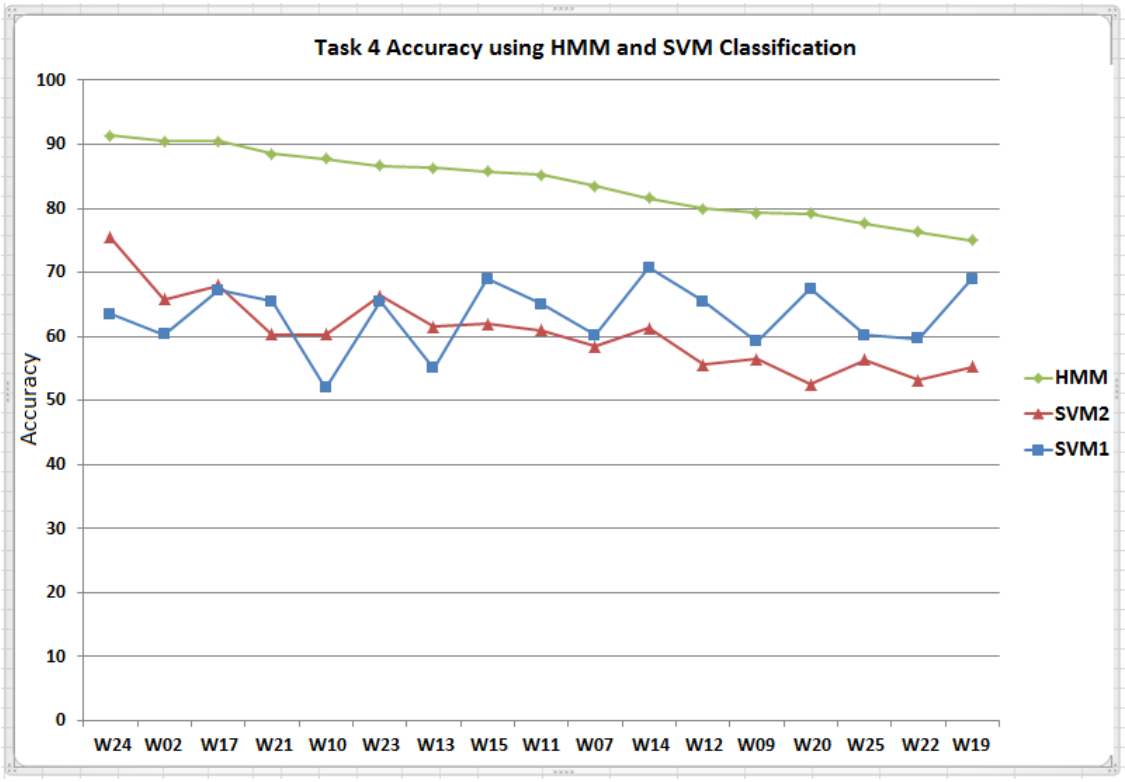

Figure 4.4: Accuracy results for the 8 consonant classes for all three classification tasks. Talkers are arranged in order of descending accuracy based on HMM results.

4.1.5 Study 1: A Comparison of Tasks 1-4

I examined the results of the 4 tasks on a per-talker basis. To accomplish this, I sorted the talkers according to the HMM accuracy results obtained in task 4 to obtain a ranking order over the talkers. I then used this ranking order to arrange the per-talker results from the other tasks, as shown in Figure 4.5.

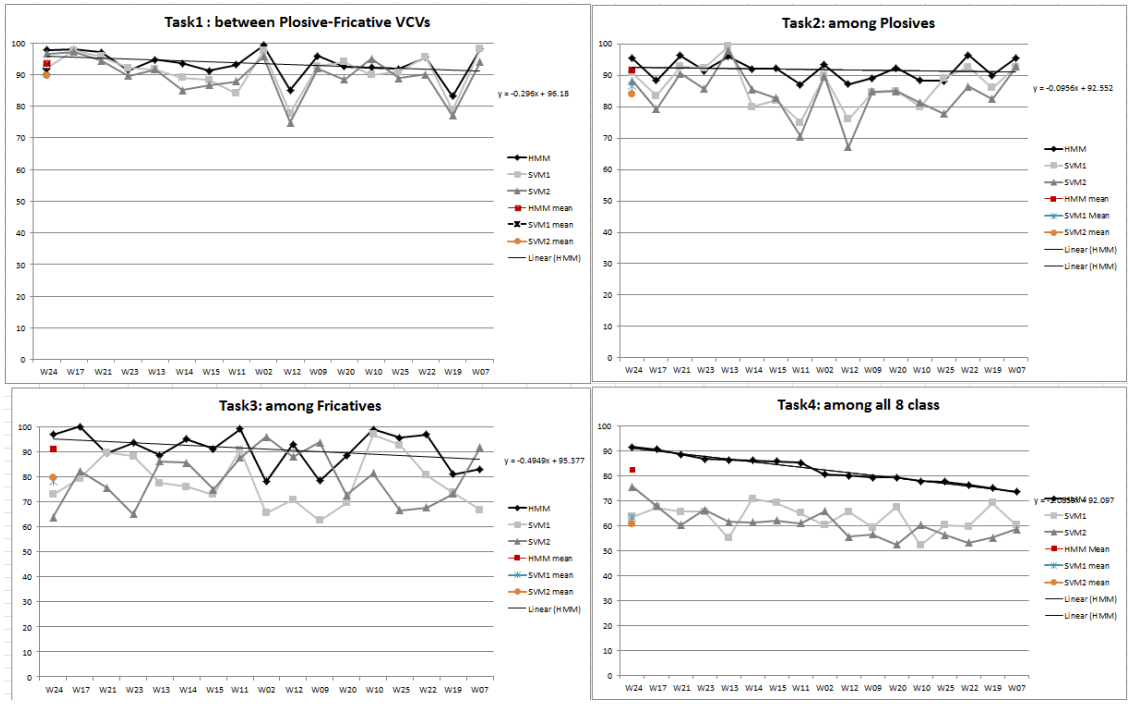

Figure 4.5: Per-talker results from classification tasks 1-4

The question at the outset was: Given a talkerfis own data concerning tongue-tip speed profiles for different consonant segments, how accurately can we classify the consonant segment type of unknown tongue-tip speed profile by that same talker? The answer to this question is in terms of technique, HMM classification technique has better mean performance over all talkers, for all four tasks.

Task 1, to distinguish between plosives and fricatives, had the best per-talker results, with 90.90%, 89.69% and 93.52% accuracy for SVM1, SVM2 and HMM, respectively. Task 4, to distinguish among all 8 consonants, had the worst per-task results, with 63.24%, 60.57% and 82.32% accuracy for SVM1, SVM2 and HMM, respectively.

4.2 Study 2 Results

Question 2: Given data representing the tongue-tip speed profiles for a pool of different talkers, how accurately can we classify the consonant segment type of unknown tongue-tip trajectories by a talker from outside that pool?

This study entailed the following four tasks:

1. Task 1: distinguish between plosives and fricatives
2. Task 2: distinguish among plosives (/t/, /d/, /k/, /g/)
3. Task 3: distinguish among fricatives (/s/, /z/, /sh/)

4. Task 4: distinguish among all 8 consonant segment classes (over three classes: plosives, fricatives, affricates)

4.2.1 Study 2, Task 1: Distinguishing between Plosives and Fricatives

This task was performed using each of two different feature sets and the SVM classification technique. The HMM classification technique used a third feature set (described in the previous chapter). In the following experiments, training was performed using a leave-one-out strategy.

4.2.1.1 Study 2, Task 1: SVM classification on the basis of Feature Set 1

The classification accuracies obtained using the SVM technique using feature set 1 are presented in table 4.11

The results were derived by considering each of the possible pairwise combinations between plosives (/t/, /d/, /g/) and fricatives (/s/, /z/, /sh/). The plosive /k/ was omitted. A mean accuracy was derived over each of these 7 possible pairings (see table 4.11).

The mean of SVM1 classification accuracy between plosive-fricative VCVs over all the talkers was 82.21%. The best mean accuracy was observed when W02 's data was used as the testing set, with 99.13% accuracy. The worst mean accuracy was observed when W23 's data as testing set was used, with 68.99% mean accuracy.

The most accurately distinguished pair was ATA-ASHA (92.36%) and the least accurately distinguished pair was AGA-AZA (73.53%).

Results of 17 folds of SVM classification as plosive or fricative VCVs using feature set 1																		
Fold	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
P-F SVM	W02 for testing	W07 for testing	W09 for testing	W10 for testing	W11 for testing	W12 for testing	W13 for testing	W14 for testing	W15 for testing	W17 for testing	W19 for testing	W20 for testing	W21 for testing	W22 for testing	W23 for testing	W24 for testing	W25 for testing	Mean
ATA-ASA	100	98.75	100	100	100	50.33	88.25	52	65.2	100	78.50	95.5	100	100	64.5	100	98.33	87.73
AZA-ATA	100	93.75	89.25	77.3	75.2	60.4	68	61.2	60.5	74.9	54.25	84.42	100	97.25	88	52.23	63.25	76.46
ATA-ASHA	100	100	100	100	100	74.50	98.23	68.3	92.12	100	88.25	90.33	100	89.25	79.2	89.9	100	92.36
ADA-ASA	95.6	60.5	100	66.2	62.4	81.33	100	87.12	94.71	90.5	90.3	100	54.2	90.82	62.3	65.2	100	82.42
ADA-AZA	98.33	66.2	100	68	70	88.33	100	96.5	66.6	58.2	80.25	78.2	62.3	82.75	60	59.2	93.75	78.15
AGA-AZA	100	80.75	80.42	63	68.50	63.15	70	100	78.75	52	79.25	69.13	67.52	80.72	61.25	56.2	79.3	73.53
ADA-ASHA	100	66.2	100	69.3	60.2	100	85.2	100	100	96.75	90.25	94.75	65.6	78.72	65.15	74.2	95.3	84.8
Mean	99.13	80.87	95.66	77.68	76.61	74.00	87.09	80.73	79.69	81.76	80.15	87.47	78.51	88.50	68.62	70.99	89.99	82.21

Table 4.11: Results of 17 folds of SVM classification as plosive or fricative VCVs using feature set 1

4.2.1.2 Study 2, Task1: SVM classification on the basis of Feature Set 2

The classification accuracies obtained using the SVM technique using feature set 2 are presented in table 4.12.

The mean of SVM2 classification accuracy between plosive-fricative VCVs over all the talkers was 77.07%. The best mean accuracy was observed when W02 's data was used as the testing set, with 99.04% accuracy. The worst mean accuracy was observed when W23 's data as testing set was used, with 59.92% mean accuracy.

The most accurately distinguished pair was ASHA-ATA (89.59%) and the least accurately distinguished pair was AGA-AZA (66.97%).

Talkers W02 and W23, whether using either feature set 1 or feature set 2, had the best and worst classification accuracy. The fricative pairs ATA-ASHA was the best distinguished, with

Results of 17 folds of SVM classification as plosive or fricative VCVs using feature set 2																		
Fold	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
P-F SVM	W02 for testing	W07 for testing	W09 for testing	W10 for testing	W11 for testing	W12 for testing	W13 for testing	W14 for testing	W15 for testing	W17 for testing	W19 for testing	W20 for testing	W21 for testing	W22 for testing	W23 for testing	W24 for testing	W25 for testing	Mean
ASA-ATA	100	93.75	100	100	100	39.33	80	40	54	100	69.50	93	100	94	55	100	93.33	83.05
AZA-ATA	100	89.75	80	70.75	65	49.93	55	49	52	65	46.75	75.42	100	93.75	76	58.33	68.66	70.31
ASHA-ATA	100	100	100	100	100	65.50	93.33	59	94.52	100	82.25	93.33	100	82.75	70	82.33	100	89.59
ADA-ASA	100	52	100	55	53	75.33	100	73.42	87.71	100	87.25	100	46.52	75.82	56	56	100	77.53
ADA-AZA	93.33	52	93.33	54	53	78.33	100	93.42	53.52	52.36	83.25	81.25	53.52	75.82	56	54	88.71	71.52
ADA-ASHA	100	52	100	65	52	93.33	100	91.42	100	93.75	81.25	100	58.52	71.72	53.25	68	88.71	80.53
AGA-AZA	100	68.75	73.42	55	64.50	63.33	63	100	71.75	45	70.25	63.33	55.52	71.72	53.25	48	71.71	66.97
Mean	99.04	72.60	92.39	71.39	69.64	66.44	84.47	72.32	73.35	79.44	74.35	86.61	73.44	80.79	59.92	66.66	87.30	77.07

Table 4.12: Results of 17 folds of SVM classification as plosive or fricative VCVs using feature set 2

92.36% and 89.59% accuracy, for each of feature set 1 and feature set 2, respectively. The fricative pairs, AGA-AZA was the least accurately distinguished pair, with 73.53% and 66.97% for each of feature set 1 and feature set 2, respectively.

4.2.1.3 Study 2, Task 1: HMM classification on the basis of Feature Set 3

HMM classification accuracies between plosive-fricative VCVs are presented in table 4.13.

The mean classification accuracies between plosive-fricative VCVs over all the talkers was 92.55%. Talkers W02, with 99.28%, and W23, with 86.15%, have the best and worst mean accuracies for consonant segment classification, respectively. The most accurately distinguished pair was ADA-ASHA (94.11%), and the least accurately distinguished pair was ATA-AZA (88.94%).

HMM classification accuracy between Plosive-fricative VCVs																		
Fold	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
P-F	W02 for testing	W07 for testing	W09 for testing	W10 for testing	W11 for testing	W12 for testing	W13 for testing	W14 for testing	W15 for testing	W17 for testing	W19 for testing	W20 for testing	W21 for testing	W22 for testing	W232 for testing	W24 for testing	W25 for testing	Mean
ATA-ASA	100	100	100	100	100	72.33	96.3	70.25	84.5	100	92.3	100	100	100	80.3	100	100	93.88
AZA-ATA	100	100	98.6	93.75	90	78.2	85.5	79.25	77.2	92.6	71	97.6	100	100	100	70.9	77.4	88.94
ATA-ASHA	100	100	100	100	100	90.3	100	86.75	100	100	100	100	100	85.2	95.6	100	91.4	97.01
ADA-ASA	100	79	100	82.5	85.3	95.2	100	100	100	88.6	100	100	78.5	100	84	89.8	98.2	93.01
ADA-AZA	100	85.2	100	81.5	88.2	100	100	96.75	87.5	75.3	95.3	92.5	80	100	79.3	76.1	100	90.45
ADA-ASHA	98.5	84.2	100	88.6	81.2	100	100	96.75	100	100	100	100	84	92.7	83.9	90	100	94.11
AGA-AZA	96.5	96.6	94.2	85.2	84.6	81.5	89	96.75	100	77.2	100	90.2	88.6	100	80	79	98.7	90.47
Mean	99.28	92.14	98.97	90.22	89.9	88.21	95.82	89.5	92.74	90.52	94.08	97.18	90.15	96.84	86.15	86.54	95.1	92.55

Table 4.13: HMM classification accuracy between Plosive-fricative VCVs

4.2.1.4 Study 2, Task 1 Summary

I next compare the Study 2, Task 1 results to examine the impact of classification technique and feature set. The data is displayed in figure 4.6.

Figure 4.6 presents the accuracies obtained using SVM (using each of feature set 1 and feature set 2) and using HMM. The mean accuracy over all the talkers and all pairing using the HMM technique was 92.55%, which is greater than SVM1 (82.21%) and SVM2 (77.07%).

4.2.2 Study 2, Task 2 Distinguishing among Plosives

This task was performed using each of two different feature sets and the SVM classification technique and the HMM classification technique using a third feature set. In this experiment, training was over a pool of all different talkers speed profiles for SVM technique and velocity

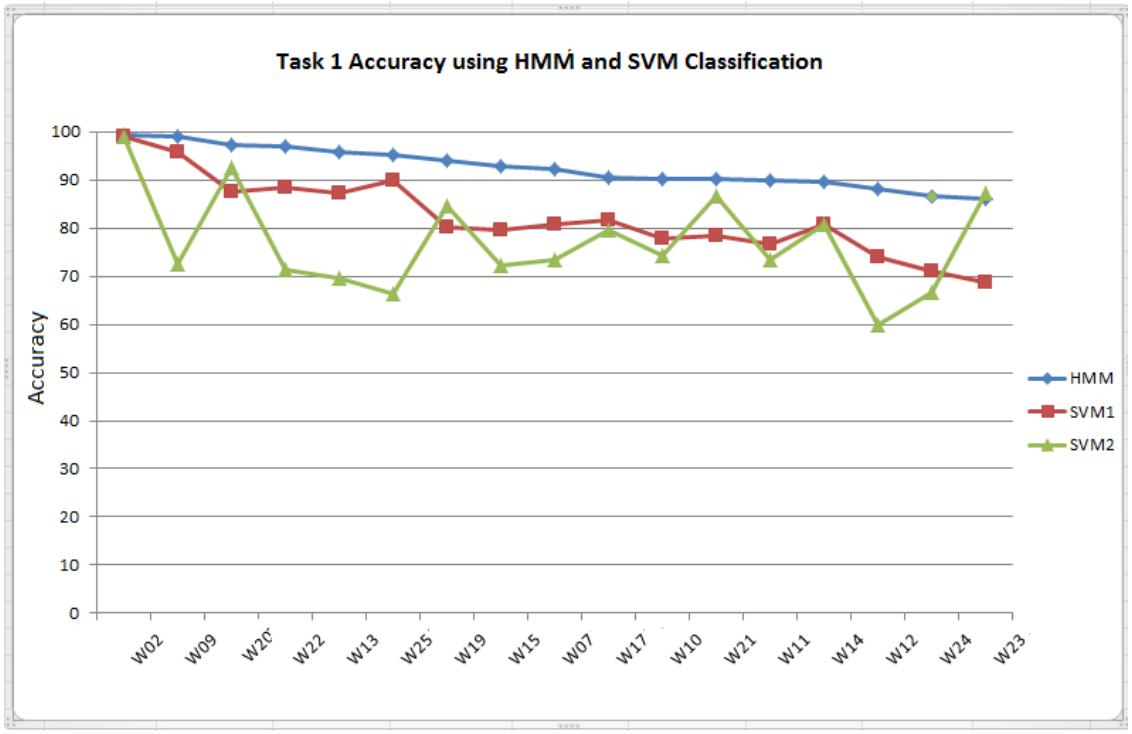

Figure 4.6: Accuracy results for plasives classification for all three classification tasks. Talkers are arranged in order of descending accuracy based on HMM results.

for HMM. training was performed using a leave-one-out strategy. The results were derived by considering each of the possible pairwise combinations among plosives (/t/, /d/, /g/). The plosive /k/ was omitted. The mean accuracy was derived over these 7 possible pairings.

4.2.2.1 Study 2, Task 2: SVM classification on the basis of Feature Set 1

The classification accuracies using feature set 1 obtained are presented in table 4.14.

Study2, Task2: SVM classification accuracies among plosives using feature 1																		
Fold	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
P-P SVM	W02 for testing	W07 for testing	W09 for testing	W10 for testing	W11 for testing	W12 for testing	W13 for testing	W14 for testing	W15 for testing	W17 for testing	W19 for testing	W20 for testing	W21 for testing	W22 for testing	W232 for testing	W24 for testing	W25 for testing	Mean
ATA-AKA	71.25	87.5	79.3	59.6	93.75	90.4	25.2	70	65.33	72.2	100	62	100	67.3	80	58.25	75	73.9
ATA-ADA	86.3	50	86.7	65.5	63.3	84.75	75	94.5	95.25	70	65.75	45.6	25	68	63	50	76.3	68.5
ATA-AGA	83	63.5	60	65.3	82.75	83.25	50.3	49.5	55.14	73.75	53.75	100	73.46	65.33	76.50	61.76	60.24	68.1
AGA-AKA	83.2	70.3	89.42	69.4	50.2	59.3	28.5	50.28	66.85	69.25	60.25	88.66	69.57	58.75	77.25	79.63	60.25	66.5
AKA-ADA	33.5	58.2	65.5	69.3	50.2	63.75	25.6	96.6	82.85	86.75	65.3	75.66	77.25	70.75	53.75	35.76	50.44	62.4
AGA-ADA	52.66	69.25	63.25	65.3	63.25	65	19	63.85	71.66	92.5	88.26	100	65.75	60.75	79.3	70	60.25	67.6
Mean	68.31	66.45	74.02	65.73	67.24	74.40	37.26	70.78	72.84	77.40	72.21	78.65	68.505	65.14	71.63	59.23	63.74	67.9

Table 4.14: Study2-task2-SVM classification accuracies among plosives using feature set 1

The mean classification accuracies among plosive VCVs over all the talkers was 67.9%. Talkers W20, with 78.65%, and W13, with 37.09%, have the best and worst mean accuracies for consonant segment classification, respectively. The most accurately distinguished pair was ATA-AKA (73.9%) and the least accurately distinguished pair was AKA-ADA (62.4%).

4.2.2.2 Study 2, Task 2: SVM classification accuracies among plosives using Feature Set 2

The classification accuracies using feature set 2 are presented in table 4.15.

SVM classification accuracy among plosives on the basis of feature set 2																		
Fold	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
P-P SVM	W02 for testing	W07 for testing	W09 for testing	W10 for testing	W11 for testing	W12 for testing	W13 for testing	W14 for testing	W15 for testing	W17 for testing	W19 for testing	W20 for testing	W21 for testing	W22 for testing	W23 for testing	W24 for testing	W25 for testing	Mean
ATA-AKA	60	75	68	50	83	81	0	60	54	60	100	50	100	58	69	46	60	63.18
ATA-ADA	75	39	75	50	50	72	60	86	86	58	53	30	12	65	60	54	68	58.41
ATA-AGA	73	52	48	53	70.75	72.75	38	35.33	42.14	62.25	40.75	93	60.57	52.25	65.50	50.66	64.42	57.32
AKA-AGA	69.66	53	72.42	52	35	45.75	10	35.28	50.85	55.75	45	75.66	55.57	43.75	64.50	63.16	55.84	51.95
AKA-ADA	10.66	43.75	49.66	52	33	48.75	12.25	85.28	65.85	67.66	55	60.66	60.25	70.75	53.75	35.76	50.44	50.32
AGA-ADA	35.66	53.25	47.66	50	50	56.75	5.25	48.85	56.66	80	75.66	100	50.75	43.75	65.50	54.33	48.25	54.25
Mean	53.99	52.66	60.12	51.16	53.62	62.83	20.91	58.45	59.25	63.94	61.56	68.22	56.52	55.58	63.04	50.65	57.82	55.90

Table 4.15: SVM classification accuracy among plosives on the basis of feature set 2

The mean classification accuracies among plosive VCVs over all the talkers was 55.9%. Talkers W20, with 68.22%, and W13, with 20.91%, have the best and worst mean accuracies for consonant segment classification, respectively. The most accurately distinguished pair was ATA-AKA (63.18%) and the least accurately distinguished pair was AKA-ADA (50.32%).

Talker W20 and W13, using either feature set 1 or feature set 2, had the best and worst classification accuracy. The plosive pair AKA-AGA was the least distinguishable, with 73.53% and 50.32% accuracy, respectively, for each of feature set 1 and feature set 2.

4.2.2.3 Study 2, Task 2: HMM classification on the basis of Feature Set 3

The HMM classification accuracies among plosive VCVs are presented in table 4.16.

HMM classification accuracies among plosives																		
Fold	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
P-P HMM	W02 for testing	W07 for testing	W09 for testing	W10 for testing	W11 for testing	W12 for testing	W13 for testing	W14 for testing	W15 for testing	W17 for testing	W19 for testing	W20 for testing	W21 for testing	W22 for testing	W23 for testing	W24 for testing	W25 for testing	Mean
ATA-AKA	86	97.25	91.2	75	100	100	46.6	85.15	82.33	89.60	100	79	100	81.21	93.75	73.5	89.33	86.47
ATA-ADA	96.5	70.6	100	85.3	87	95.25	85.6	100	100	75	80.25	62.75	53.5	83.3	78.6	69.9	92.2	83.28
ATA-AGA	95.5	78.3	75.3	81.75	94.2	90.3	71.2	70.75	78.32	85.3	70.3	100	93.5	80.25	100	79.75	74.33	83.47
AGA-AKA	96.6	87.3	100	85.3	68.9	75.6	55	69.75	83.2	87	85.2	77.25	88.26	72.45	100	100	81.4	83.13
AKA-ADA	55.2	78.3	84.3	75.6	80.33	55.33	100	96.6	100	84.25	90	80	92.5	86.6	70.2	60	71.3	80.03
AGA-ADA	70.2	85.3	80	83	81.5	84.25	45	79.75	88.3	100	100	100	82.25	79	96.75	86.3	79.25	83.58
Mean	83.33	82.84	88.46	80.99	85.32	83.45	67.23	83.66	88.69	86.85	87.62	83.8	85.001	80.46	89.88	78.24	81.30	83.36

Table 4.16: HMM classification accuracies among plosives

The mean classification accuracies among plosive VCVs over all the talkers was 83.36%. Talkers W23, with 89.88%, and W13, with 67.23%, have the best and worst mean accuracies for consonant segment classification, respectively. The most accurately distinguished pair was ATA-AKA (86.47%) and the least accurately distinguished pair was AKA-ADA (80.03%).

4.2.2.4 Study 2, Task 2 Summary

I next compare the Study 2, Task 2 results to examine the impact of classification technique and feature set. The data is displayed in figure 4.7.

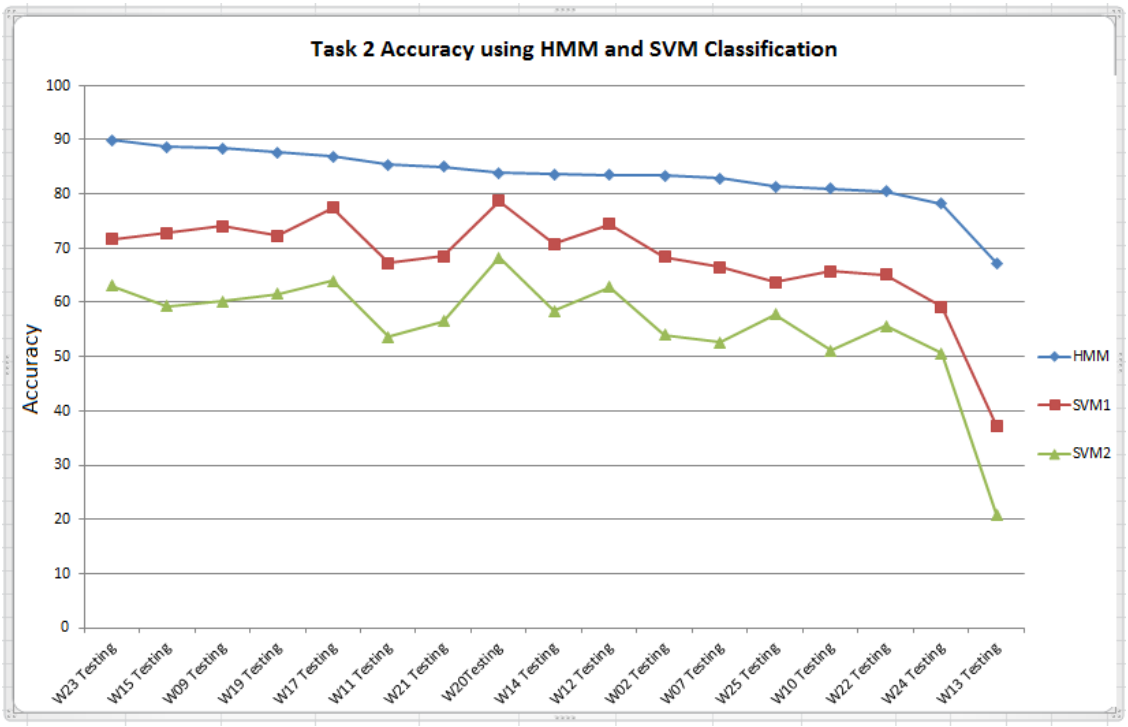

Figure 4.7: Accuracy results for plosives classification for all three classification tasks. Talkers are arranged in order of descending accuracy based on HMM results.

The mean accuracy over all the talkers and all pairing using the HMM technique was 83.36%, which is greater than SVM1 (67.60%) and SVM2 (55.90%).

4.2.3 Study 2, Task 3: Distinguishing among Fricatives

This task was performed using each of two different feature sets and the SVM classification technique. The HMM classification technique used a third feature set. In this experiment, training was performed over a pool of all different talkers'speed profiles for SVM technique and velocity profiles for HMM, using a leave-one-out strategy. The results were derived by considering each of the possible pairwise combinations between fricatives (/s/, /z/, /sh/). The mean accuracy was derived over these 3 possible pairings.

4.2.3.1 Study 2, Task 3: SVM Classification on the basis of Feature Set 1

SVM classification accuracies obtained using feature set 1 to distinguish among fricative VCVs are presented in table 4.17.

SVM classification accuracy among fricatives using feature set 1																		
Fold	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
F-F SVM	W02 for testing	W07 for testing	W09 for testing	W10 for testing	W11 for testing	W12 for testing	W13 for testing	W14 for testing	W15 for testing	W17 for testing	W19 for testing	W20 for testing	W21 for testing	W22 for testing	W23 for testing	W24 for testing	W25 for testing	Mean
AZA-ASA	44.66	59.15	55.6	59	55	64.5	14.15	55.85	67.3	85	84.2	93.25	62.75	54.15	78.50	69.13	60.75	62.52
ASA-ASHA	100	62.66	58.16	100	84	75.15	53.75	94.25	56.66	93.2	46.26	91.2	59.75	65.15	74.3	96.5	85.6	76.27
AZA-ASHA	54.3	63.4	96.66	100	75	77.75	83.15	94.5	100	100	46.26	78.2	68.75	57.75	73.50	65.5	90.15	77.93
Mean	66.32	61.73	70.14	86.33	71.33	72.46	50.35	81.53	74.65	92.73	58.90	87.55	63.75	59.01	75.43	77.043	78.83	72.24

Table 4.17: SVM classification accuracy among fricatives using feature set 1

The mean classification accuracies among fricative VCVs over all the talkers was 72.24%. Talkers W17, with 92.73%, and W13, with 50.35%, have the best and worst mean accuracies for consonant segment classification, respectively. The most accurately distinguished pair was AZA-ASHA (77.93%). The least accurately distinguished pair was AZA-ASA (62.52%).

4.2.3.2 Study 2, Task 3: SVM Classification on the basis of Feature Set 2

SVM classification accuracies using feature set 2 are presented in table 4.18.

SVM classification accuracy among fricatives using feature set 2																		
Fold	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
F-F SVM	W02 for testing	W07 for testing	W09 for testing	W10 for testing	W11 for testing	W12 for testing	W13 for testing	W14 for testing	W15 for testing	W17 for testing	W19 for testing	W20 for testing	W21 for testing	W22 for testing	W232 for testing	W24 for testing	W25 for testing	Mean
ASA-AZA	35.66	53.25	47.66	50	50	56.75	5.25	48.85	56.66	80	75.66	100	50.75	43.75	65.50	54.33	48.25	55.42
ASA-ASHA	100	50.66	50.66	100	78	66.75	45.75	88.84	48.66	100	38.26	80.85	50.75	53.75	68.50	90.33	88.25	71.33
AZA-ASHA	45.5	52.66	60.66	100	68	68.75	75.75	88.84	95.66	100	38.26	70.85	55.75	46.75	66.50	58.35	80.25	70.29
Mean	60.38	59.67	52.99	83.33	65.33	64.08	53.13	75.51	66.99	93.33	50.72	83.9	52.41	48.08	66.83	67.67	72.25	65.68

Table 4.18: SVM classification accuracy among fricatives using feature set 2

The mean classification accuracies among fricative VCVs over all the talkers was 65.68%. Talkers W17, with 93.33%, and W22, with 48.08%, had the best and worst mean accuracies for consonant segment classification, respectively. The most accurately distinguished pair was ASA-ASHA (71.33%). The least accurately distinguished pair was ASA-AZA (55.42%).

Distinguishing between fricatives AZA and ASA, on the basis of either feature set 1 or feature set 2, had the least accuracy over all the possible fricative pairings, with 62.52% and 55.42% accuracy, respectively. W17 on the basis of either feature set 1 or feature set 2, had the best

mean accuracy over all the talkers, with 92.73% and 93.33%, respectively,

4.2.3.3 Study 2, Task 3: HMM classification using Feature Set 3

HMM classification accuracies in distinguishing among fricative VCVs are presented in table 4.19.

HMM classification accuracies among fricatives																		
Fold	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
F-F HMM	W02 for testing	W07 for testing	W09 for testing	W10 for testing	W11 for testing	W12 for testing	W13 for testing	W14 for testing	W15 for testing	W17 for testing	W19 for testing	W20 for testing	W21 for testing	W22 for testing	W23 for testing	W24 for testing	W25 for testing	Mean
AZA-ASA	69.5	77.3	76	79	76.25	82.2	40	79.3	86.5	100	100	100	85.5	77.2	94.25	90.75	82.3	82.12
ASA-ASHA	100	82.5	75.6	100	100	94.2	72.9	100	78.5	100	67.2	100	80.5	86.3	93.75	100	100	90.08
AZA-ASHA	76	85.2	100	100	65.3	92.5	97.2	100	100	100	68.5	95.4	89.2	77.25	92.7	88.6	100	89.87
Mean	81.83	81.66	83.86	93	80.516	89.63	70.03	93.1	88.33	100	78.56	98.46	85.06	80.25	93.56	93.11	94.1	87.35

Table 4.19: HMM classification accuracies among fricatives

The most accurately distinguished pair was ASA-ASHA (90.08%) and the least accurately distinguished pair was AZA-ASA (82.12%).

4.2.3.4 Study 2, Task 3 Summary

I next compared the Study 2, Task 3 results to examine the impact of classification technique and feature set. The data is displayed in figure 4.8.

The mean accuracy over all the talkers and all pairing using the HMM technique was 87.35%, which is greater than SVM1 (72.24%) and SVM2 (65.68%). SVM classification accuracy using feature set 1 would seem to have better classification accuracy than using feature set 2.

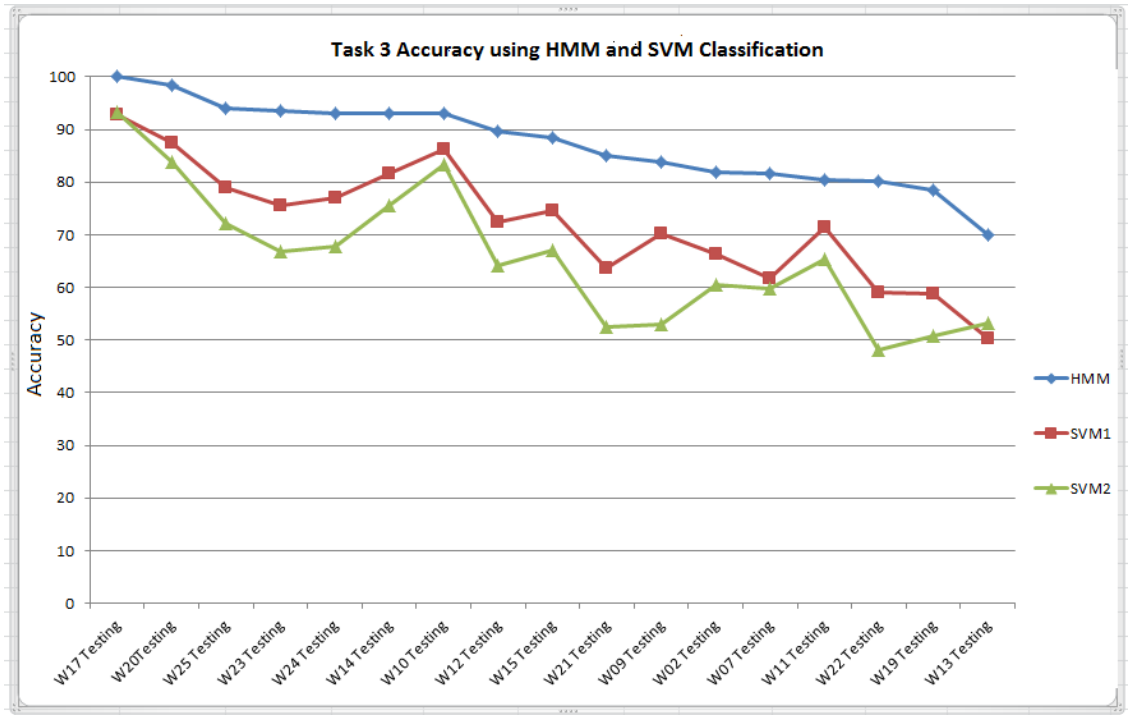

Figure 4.8: Accuracy results for fricatives classification for all three classification tasks. Talkers are arranged in order of descending accuracy based on HMM results.

4.2.4 Study 2, Task 4 Distinguishing among all 8 consonants

The SVM1, SVM2, and HMM results on a per talker basis are provided in table 4.20.

Classification accuracies for each talker, for each of three different classification techniques (SVM1, SVM2, and HMM).																		
Fold	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Speaker for testing	W25 for testing	W24 for testing	W23 for testing	W22 for testing	W21 for testing	W20 for testing	W19 for testing	W17 for testing	W15 for testing	W14 for testing	W13 for testing	W12 for testing	W11 for testing	W10 for testing	W09 for testing	W07 for testing	W02 for testing	Mean
SVM1 Acc	62.27	63.41	61.61	60.85	60.95	59.57	59.92	58.26	60.63	62.19	62.39	66.05	62.55	58.92	57.35	46.31	53.88	59.83
SVM2 Acc	60.41	57.77	57.92	58.11	58.19	60.42	57.76	60.27	56.89	58.65	56.62	60.39	59.46	59.31	51.41	56.14	65	58.51
HMM Acc	80.26	82.6	85	77.3	83.95	86.57	82.92	80.26	81.63	88.19	83.39	79.05	76.55	88.92	82.35	70.31	77.88	81.59

Table 4.20: Classification accuracies in distinguishing among all 8 consonants, for each talker, for each of three different classification techniques (SVM1, SVM2, and HMM).

The mean accuracies over all talkers for this task were 81.59%, 59.83% and 58.51%, for HMM, SVM1 and SVM2 respectively. The talkers that had the best classification accuracies were W12 with 66.05%, W02 with 65%, and W10 with 88.92% for SVM1, SVM2 and HMM, respectively.

I next examined the results on a per-VCV basis, as shown in table 4.21.

HMM classification results distinguish among all 8 classes, True positive Rate	
$phrase/Rate$	$TPRate$
ADA	0.72
AGA	0.654
AKA	0.703
ASA	0.79
ATA	0.937
AZA	0.505
ASHA	0.859
ATCHA	0.968
Average	0.812

Table 4.21: HMM classification true positive rate among all 8 consonant classes.

The segments ATCHA, with 96%, and AZA, with 50%, are those with the highest and lowest classification true positive rates, respectively.

4.2.5 Study 2, Task 4 Summary

I next compare the Study 2, Task 4 results to examine the impact of classification technique and feature set. The results based on using a leave-one-out strategy over all the talkers are displayed in figure 4.9.

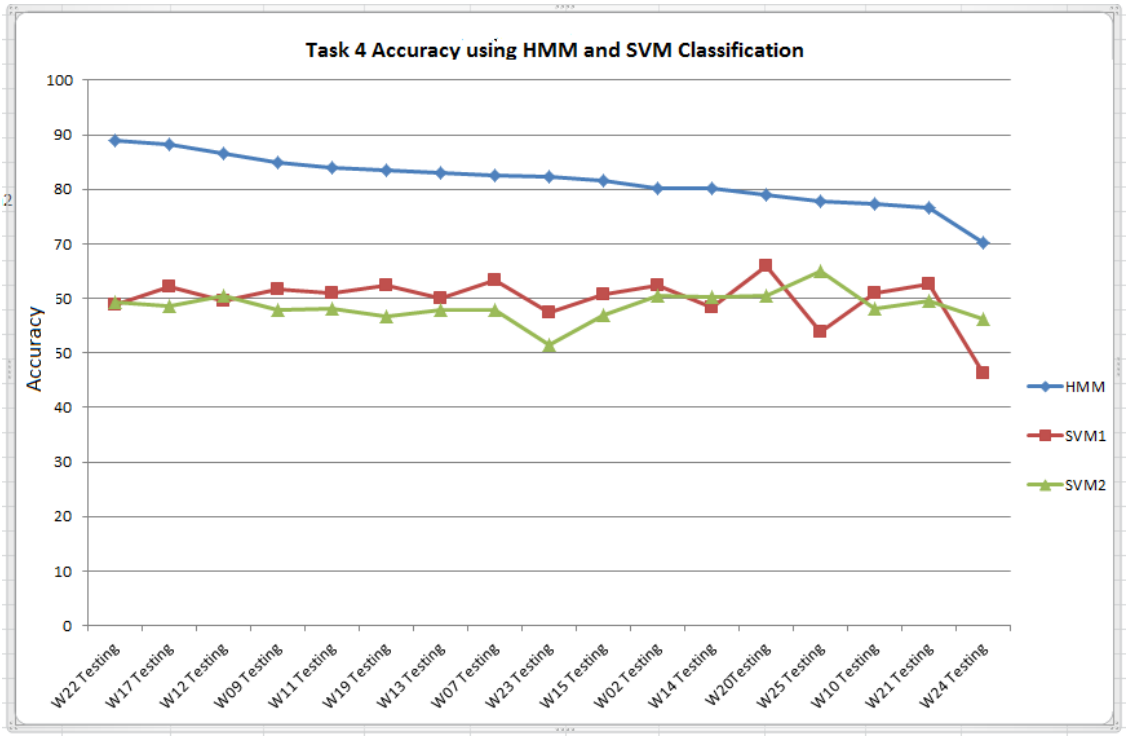

Figure 4.9: Accuracy results for all 8 consonants classification for all three classification tasks. Talkers are arranged in order of descending accuracy based on HMM results.

The mean accuracy over all the talkers and all pairing using the HMM technique was 81.59%,

which is greater than SVM1 (59.83%) and SVM2 (58.51%). SVM classification accuracy using feature set 1 would seem to have better classification accuracy than using feature set 2.

4.2.5.1 Study 2, Comparison over Tasks 1-4

I examined the results of the 4 tasks of study 2 on a per-talker basis. To accomplish this, I sorted the talkers according to the HMM accuracy results obtained in task 4 to obtain a ranking order. I then used that ranking order to arrange the talkers to present the results from the other tasks. The results are presented in figure 4.10.

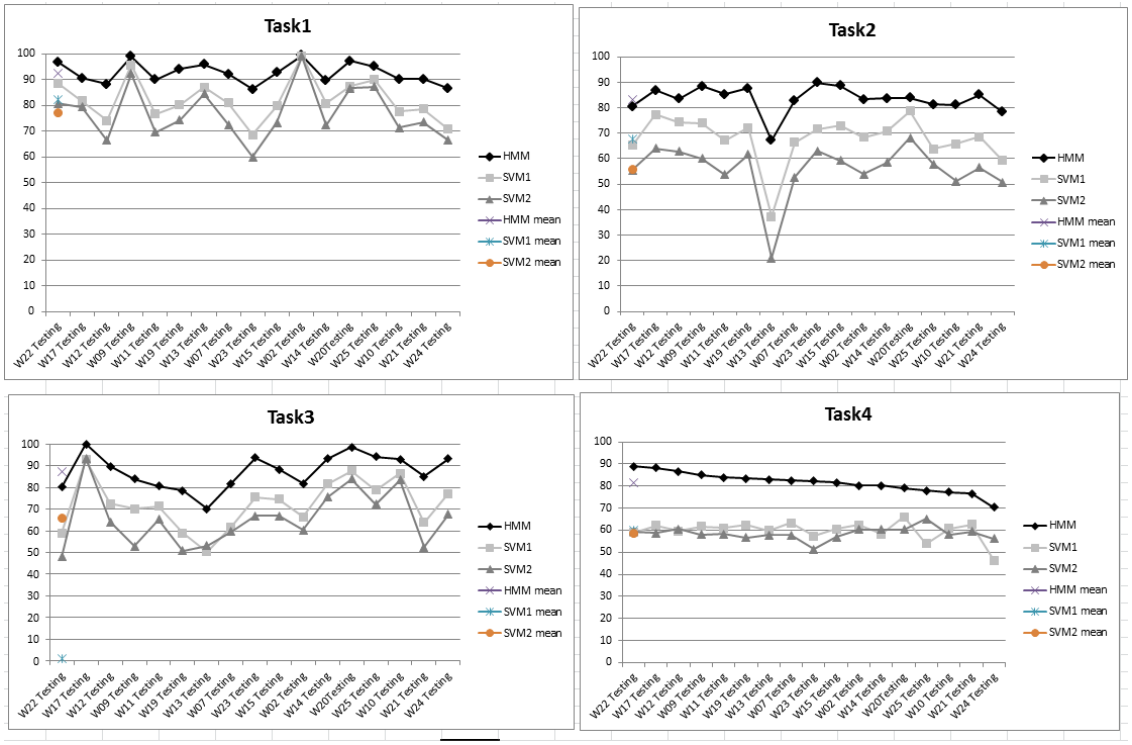

Figure 4.10: Results of all of the tasks within study 2, sorted in order on the based on task 4 HMM accuracies

The question at the outset was: Given data representing the tongue-tip speed profiles for a pool of different talkers, how accurately can we classify the consonant segment type of unknown tongue-tip trajectories by a talker from outside that pool?

The answer to this question is, in terms of technique, HMM classification technique has better performance over SVM1 and SVM2 for all four tasks.

Over all the SVM1, SVM2 and HMM, the best classification accuracies were observed for task 1, distinguishing between plosive-fricative VCVs, with 82.20%, 77.07% and 92.55% mean accuracies for SVM1, SVM2 and HMM, respectively. The worst classification accuracies were observed for Task 4 with 59.83%, 58.51% and 81.59% mean accuracies, for SVM1, SVM2 and HMM, respectively.

4.3 Study 1 and Study 2, A Comparison of Training Techniques

I next examined the talkers, looking for differences in results between study 1 and study 2, in order to gain insight into the impact of training technique.

One group of subjects, including W02, W10, W21, W24, obtained relatively high mean classification accuracies in study one, but then obtained the relatively low mean accuracies in study 2. They obtained better results in per-talker study than results which is obtained in study 2, using a leave-one-out strategy. This implies that their classification of their own segments is best accomplished when training using their own data, as opposed to training using the data from other talkers.

Another group of talkers, including W7, W11, W13, W14, W15, W17, W20, W23, W25,

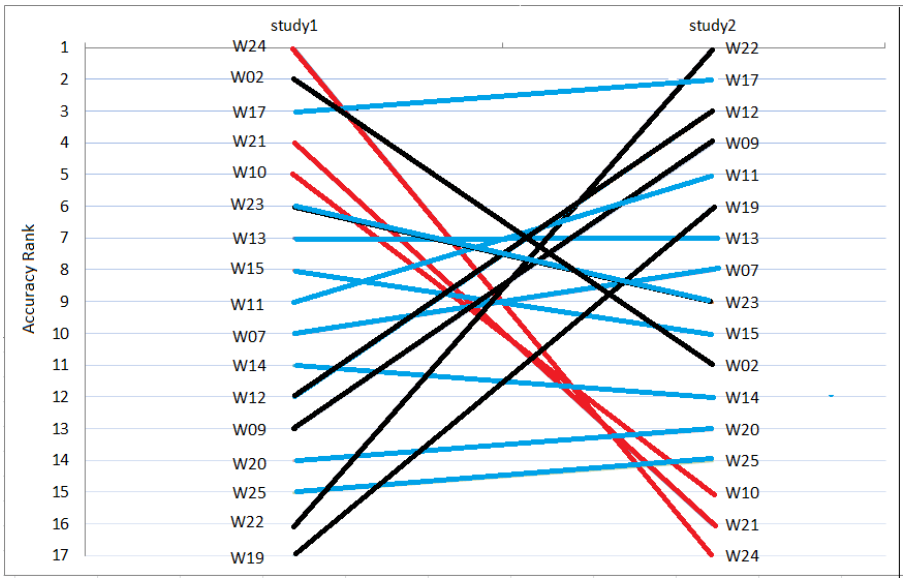

Figure 4.11: Impact of feature set: differences between study 1 and study 2.

obtained results in study 1 and study 2 that were not much different. we can infer that training using the talker's own or other talkers'data yielded similar results.

The third and final group of talkers, including W9, W12, W19, W22, obtained relatively low mean classification accuracy in study one, but relatively high mean accuracies in study 2. For these talkers, training using their own data yielded worse results than training using the data of other talkers. These talkerfis own datasets possessed less information than those of others that had utility to the classifier.

4.4 Study 3: Results

Question 3: Given data representing the tongue-tip speed profiles for a pool of different talkers, how accurately can we classify the consonant segment of unknown tongue-tip trajectories by a

talker from within that pool?

4.4.1 Study 3, Task 1: Distinguishing between Plosive-Fricative VCVs

This task was performed using each of two different feature sets and the SVM classi- fication technique. The HMM classification technique used a third feature set. In this experiment, training was performed over the pooled data of all the talkers'speed profiles for SVM technique and the velocity profiles for HMM. The task was to distinguish between Plosive-Fricative VCVs. Testing was performed using one data set. The same type of datasets were used for each task (for SVM and HMM respectively, except all talker data was pooled together and training/testing splits used). Folds were drawn over all speakers and I used four-fold cross-validation. The training set was composed of 75% of the data which reflects 75% of each speakerfis repetitions. Testing was performed using the other 25% of the repetitions.

The mean classification accuracies obtained using SVM1, SVM2 and HMM classification techniques are presented in table 4.22.

The mean accuracy over all different plosives and fricatives consonant pairs using the HMM technique was 92.8%, which is greater than SVM1 (81.47%) and SVM2 (75.07%). SVM classification accuracy using feature set 1 would seem to have better classification accuracy than using feature set 2.

SVM1, SVM2 and HMM classification accuracy between Plosive-fricative consonants			
classifier	SVM1	SVM2	HMM
ATA-ASA	94	85	100
ATA-AZA	93	84	100
ATA-ASHA	94	85	100
ADA-ASA	73.25	75.25	92.5
ADA-AZA	65	63.2	85.3
ADA-ASHA	67.5	62.7	80
AGA-ASA	84	73.7	89.2
AGA-AZA	76.5	76.5	89
AGA-ASHA	86	70.2	100
Mean	81.47	75.07	92.8

Table 4.22: SVM1, SVM2 and HMM classification accuracy between Plosive-fricative consonants

The plosive- fricative pair, ATA-ASA, had the best distinguishability, with 94%, 85%, and 100% mean accuracy over all talkers, for SVM1, SVM2 and HMM, respectively. The plosive- fricative pair, ADA-ASHA had the lowest distinguiability, with 67.5%, 62.7% and 80% mean accuracy over all talkers, for SVM1, SVM2 and HMM, respectively.

4.4.2 Study 3, Task 2: Distinguishing among Plosives

The classifiers were retained to distinguish among the plosives. The mean classification accuracies obtained using SVM1, SVM2 and HMM classification techniques are presented in table 4.23.

The mean accuracies over all different plosive consonant pairs using the HMM technique was 94.11%, which is greater than SVM1 (78.25%) and SVM2 (68.3%).

For SVM1 and HMM, the pair that demonstrated the best classification accuracy, was ATA-AGA, with 92.75% and 100% respectively. For SVM2, the pair that demonstrated the best clas-

HMM, SVM1 and SVM2 mean accuracies among plosives			
classifier	SVM1	SVM2	HMM
ATA-ADA	86	66.2	97
ATA-AKA	92	65	88
ATA-AGA	92.75	65.7	100
ADA-AKA	85.75	69	100
ADA-AGA	74.75	70.2	93.5
AKA-AGA	70	63.2	83.2
mean	78.25	68.3	94.11

Table 4.23: HMM, SVM1 and SVM2 mean accuracies among plosives

sification accuracy was ADA-AGA with 70.2%. The plosive- fricative pair, AKA-AGA was the least accurately distinguished pair, with 70%, 63.2% and 83.2% for SVM1, SVM2 and HMM, respectively.

4.4.3 Study 3, Task 3: Distinguishing among Fricatives

I examined the classification results obtained using SVM1, SVM2 and HMM classification techniques to distinguish among fricative consonants. The mean accuracies are presented in table 4.24.

HMM, SVM1 and SVM2 overall Accuracy among fricatives			
classifier	SVM1	SVM2	HMM
ASA-AZA	73	65	86.5
ASA-ASHA	75	71	89
AZA-ASHA	67	67	90
Mean	71.6	67.6	88.3

Table 4.24: HMM, SVM1 and SVM2 mean accuracies among fricatives

The mean accuracies over all different fricative consonant pairs using the HMM technique was 88.3%, which is greater than SVM1 (71.6%) and SVM2 (67.6%).

For SVM1 and SVM2, the pair that demonstrated the best classification accuracy, was ASA-ASHA, with 75% and 71% respectively. For HMM, the pair that demonstrated the best classification accuracy was AZA-ASHA with 90%. The plosive-fricative pair, ASA-AZA was the least accurately distinguished pair, with 65%, 86.5% for SVM2 and HMM, respectively. For SVM1, the pair that demonstrated the least classification accuracy was AZA-ASHA with 67%.

4.4.4 Study 3, Task 4: Distinguishing among all 8 consonants

I examined the classification results obtained using SVM1, SVM2 and HMM classification techniques to distinguish among all 8 consonants.

4.4.4.1 Study 3, Task 4: SVM classification on the basis of Feature Set 1

In terms of the statistical used techniques, folds are drawn over all speakers. Testing folds include only a subset of repetitions per speaker and all speakers are represented in each fold. Using four fold matrices the overall confusion matrix is also calculated. The mean of correct classification (Acc) is 68.85. The following illustrates confusion matrix which are displaying the results over all the four fold cross validation in table 4.25.

SVM classification confusion matrix distinguishing among all 8 classes feature set 1								
Phrase	*ASA*	*ATA*	*AKA*	*AZA*	*AGA*	*ASHA*	*ATCHA*	*ADA*
ASA	0.6880	0	0	0.1188	0.0113	0.2025	0.0638	0.0223
ATA	0	0.6560	0.1035	0.0528	0.0113	0.0113	0.1350	0.1350
AKA	0.0185	0	0.6149	0.0838	0.0838	0	0.0454	0.1223
AZA	0.1135	0.1890	0.0445	0.5070	0.0123	0.1135	0.0445	0.0123
AGA	0	0.0345	0.1459	0.1459	0.5350	0	0.0690	0.0345
ASHA	0.0613	0.0223	0	0.1290	0	0.6510	0.1158	0.0645
ATCHA	0.1015	0.0612	0.0303	0.1212	0.0603	0.0909	0.742	0
ADA	0	0.2081	0.0445	0.1290	0.1013	0.0323	0.0468	0.6230
Mean: 68.85								

Table 4.25: SVM classification confusion matrix distinguishing among all 8 classes

4.4.4.2 Study 3, Task 4: SVM classification on the basis of Feature Set 2

Using four fold matrices the overall confusion matrix is also calculated. The mean of correct classification (Acc) is 64.76. The following illustrates confusion matrix which are displaying the results over all the four fold cross validation in table 4.26.

SVM classification confusion matrix distinguish among all 8 classes feature set 2								
Phrase	*ASA*	*ATA*	*AKA*	*AZA*	*AGA*	*ASHA*	*ATCHA*	*ADA*
ASA	0.7882	0.0882	0	0.1086	0	0.2025	0.0638	0.0325
ATA	0	0.6635	0.1563	0.0938	0.0313	0.0313	0.1875	0.1875
AKA	0.0385	0	0.5801	0.1538	0.1538	0	0.1154	0.1923
AZA	0.1935	0.1290	0.0645	0.4691	0.0323	0.1935	0.0645	0.0323
AGA	0	0.0345	0.2759	0.2759	0.5102	0	0.0690	0.0345
ASHA	0.0613	0	0	0.1290	0	0.6840	0.2258	0.0645
ATCHA	0.1515	0.1212	0.0303	0.1212	0.0603	0.0909	0.7392	0
ADA	0	0.2581	0.0645	0.1290	0.1613	0.0323	0.0668	0.6472
Mean: 64.76								

Table 4.26: SVM classification confusion matrix distinguish among all 8 classes feature set 2

4.4.4.3 Study 3, Task 4: HMM classification using Feature Set 3

I experimented with a number of different cross validation folds and states to examine the impact of these parameters on the classification accuracies in distinguishing among all 8 classes using the HMM classifier. I determined the the best classification accuracies were obtained for the following parameter values:

1. number of states = 41
2. cross validation = 22 folds
3. iterationCutoff = 0.01

The following illustrates confusion matrix which are displaying the HMM classification results distinguishing among all 8 classes over all the four fold cross validation in table 4.27.

HMM classification results distinguishing among all 8 classes								
$Phrase$	ASA	ATA	AKA	AZA	AGA	$ASHA$	$ATCHA$	ADA
ADA	79	2.2	1.5	1.5	11.19	2.2	1.5	0.7
AGA	0.72	89.05	4.3	0	3.64	1.45	0.72	0
AKA	4.2	11.42	77.85	1.42	3.57	0	0	1.42
ASA	1.42	0.71	2.14	82.14	2.85	6.42	0.71	3.57
ATA	6.38	2.12	1.41	2.12	82.97	0.70	0.70	3.54
AZA	3.64	2.18	1.45	10.94	2.18	76.64	0.72	2.18
ASHA	0.78	0	0.78	1.57	3.93	3.14	0.84	5.51
ATCHA	3.05	0	0	0	7.63	1.52	4.58	83.20

Table 4.27: HMM confusion matrix, classification results distinguishing among all 8 classes

The segments with the the highest and lowest classification accuracies were AGA with 95% and AZA with 84% have the highest and lowest true positive results in distinguishing among all

HMM classification true positive and false positive results in distinguishing among all 8 classes.		
$phrase/Rate$	$TPRate$	$FPRate$
ADA	0.874	0.32
AGA	0.955	0.030
AKA	0.842	0.018
ASA	0.908	0.028
ATA	0.917	0.054
AZA	0.848	0.024
ASHA	0.929	0.014
ATCHA	0.918	0.026
Mean	0.883	0.019

Table 4.28: HMM classification true positive and false positive results in distinguishing among all 8 classes.

8 classes.

The mean true positive rate of HMM is greater than SVM1 and SVM2. The results demonstrated 68.85%, 64.76% and 88.5% the mean true positive over all eight classes for SVM1, SVM2 and HMM classification technique.

ASA-ASHA in both SVM1 and SVM2 with 75% and 71% and AZA-ASHA in HMM with 90% demonstrated the best true positive results.

4.4.4.4 Study 3, Comparison over Tasks 1-4

The question at the outset was: Given data representing the tongue-tip speed profiles for a pool of different talkers, how accurately can we classify the consonant segment type of unknown tongue-tip trajectories by a talker from within that pool?

The answer to this question is, in terms of technique, HMM classification technique had

better performance over all four tasks. Over all the SVM1, SVM2 and HMM, task1, distinguishing between plosive-fricative VCVs results with 82.20%, 77.07% and 92.55% mean accuracy for SVM1, SVM2 and HMM, have performed better than other tasks, respectively. Task 4, distinguishing among all 8 consonants with 59.83%, 58.51% and 81.59% for SVM1, SVM2 and HMM, respectively, was the least distinguishable.

4.5 HMM Classification in Study 1 and Study 2

The accuracy of the HMM classification technique in study 1 and study 2 was significantly different, ($p<0.05$, $t(8) = 2.675370$). Classification of the consonant segment type of unknown tongue-tip trajectories by a talker from within the pool had better results than classification of the consonant segment type of unknown tongue-tip trajectories by a talker from outside the pool.

4.6 Conclusion

In terms of technique, HMM classification technique had better performance over all three studies to classify the consonant segment type of unknown tongue-tip speed profile by the same talker, by a talker from outside the pool, and by a talker from within the pool.

Over all three studies, the best classification accuracies were observed for task 1, distinguishing between plosive-fricative VCVs results for SVM1, SVM2 and HMM.

Over all three studies, the worst classification accuracies were observed for Task 4, distinguishing among all 8 consonants for SVM1, SVM2 and HMM.

The difference between the accuracies obtained using the HMM classification technique in study one (81.47%) and the HMM techniques in study two (88.3%) among all consonants was determined to be significant. Classification of the consonant segment type of unknown tongue-tip trajectories by a talker from within the pool had better results than classification of the consonant segment type of unknown tongue-tip trajectories by a talker from outside the pool.

Chapter 5

Conclusion and Future work

5.1 Findings

A long term goal of this research is to develop a new game-based speech therapies for articulatory disorders. These systems will require knowledge about the trajectories of the tongue tip during the articulation of different speech sounds and about how to best distinguish among them in order to produce helpful feedback to the user. These CBST systems will make use of kinematic signals for speech therapy interventions for different consonant segments.

In Chapter 2, I provided a literature review of speech motor control, Electromagnetic Articulography (EMA) technology, and Computer-Based Speech Therapy (CBST) systems and clinical targets. This review supports the need for classification-based approaches for kinematic speech data for clinical targets. Whereas several prior computer-based approaches have focused on the use of clinical objectives that concern spatialized aspects of the tongue-tip trajectory (e.g., the targeting of improved accuracy in lingual-palate contact for certain phonemic segments), this line of inquiry focuses on the potential use of attributes relating to the speed of the tongue-tip

trajectory as an alternative clinical objective. We situate our work in the body of prior work on the velocity characteristics of different phonemic segments. For speed-based clinical targets to be viable, however, it is necessary to characterize and to analyze the relative amounts of variability among and within talkers and phonemic segments with respect to speed-related characteristics. As an intermediate goal, I have analyze tongue-tip trajectories and determined the speed and velocity characteristics of a set of 8 different phonemes (8 different consonants, (/D/, /K/, /S/, /T/, /Z/, /G/, /SH/, /TCH/), which were obtained using the WAVE electromagnetic articulography system. I sought to classify the trajectories using two classification approaches, HMM and SVM. I structured this work around the following three questions:

1. Qyestion 1: Given a talkerfis own data concerning tongue-tip kinematic profiles for different consonant segments, how accurately can we classify the consonant segment of unknown tongue-tip kinematic profiles by that same talker?

2. Qyestion 2: Given data representing the tongue-tip kinematic profiles for a pool of different talkers, how accurately can we classify the consonant segment of unknown tongue-tip kinematic profiles by a talker from outside that pool?

3. Qyestion 3: Given data representing the tongue-tip kinematic profiles for a pool of different talkers, how accurately can we classify the consonant segment of unknown tongue-tip kinematic profiles by a talker from within that pool?

Chapter 3 described the design space, the development process, and the requirements, and evaluation strategies. I described the suite of three validation studies that I designed in or-

der to answer these research questions, and the data collection and preparation procedure for each study. I described the approaches which focuses on a large kinematic speech dataset that includes multiple repetitions of 8 different phonemic segments (/d/, /k/, /s/, /t/, /z/, /g/, /sh/, /tch/) by 18 talkers. The study design entails the use of both SVM and HMM classification, and each study makes use of datasets that employ different sets of feature, which have been derived from the speed and velocity properties of the kinematic data. Chapter 4 presented the results of the three studies (four tasks for each, concerning analysis of HMM and SVM classification approaches).

Study 1 provides the results of the 4 tasks on a per-talker basis. To accomplish this, I sorted the talkers according to the HMM accuracy results obtained in task 4 to obtain a ranking order over the talkers. I then used this ranking order to arrange the per-talker results from the other tasks. The question at the outset was: Given a talker's own data concerning tongue-tip speed profiles for different consonant segments, how accurately can we classify the consonant segment type of unknown tongue-tip speed profile by that same talker? The answer to this question is in terms of technique, HMM classification technique has better mean performance over all talkers, for all four tasks. Task 1, to distinguish between plosives and fricatives, had the best per-talker results, with 90.90%, 89.69% and 93.52% accuracy for SVM1, SVM2 and HMM, respectively. Task 4, to distinguish among all 8 consonants, had the worst per-task results, with 63.24%, 60.57% and 82.32% accuracy for SVM1, SVM2 and HMM, respectively.

Study 2 presented the results of the 4 tasks on a per-talker basis. To accomplish this, I sorted the talkers according to the HMM accuracy results obtained in task 4 to obtain a ranking

order. I then used that ranking order to arrange the talkers to present the results from the other tasks. The question at the outset was: Given data representing the tongue-tip speed profiles for a pool of different talkers, how accurately can we classify the consonant segment type of unknown tongue-tip trajectories by a talker from outside that pool? The answer to this question is, in terms of technique, HMM classification technique has better performance over SVM1 and SVM2 for all four tasks. Over all the SVM1, SVM2 and HMM, the best classification accuracies were observed for task 1, distinguishing between plosive-fricative VCVs, with 82.20%, 77.07% and 92.55% mean accuracies for SVM1, SVM2 and HMM, respectively. The worst classification accuracies were observed for Task 4 with 59.83%, 58.51% and 81.59% mean accuracies, for SVM1, SVM2 and HMM, respectively.

The study 3 question was: Given data representing the tongue-tip speed profiles for a pool of different talkers, how accurately can we classify the consonant segment type of unknown tongue-tip trajectories by a talker from within that pool? The answer to this question is, in terms of technique, HMM classification technique had better performance over all four tasks. Over all the SVM1, SVM2 and HMM, task1, distinguishing between plosive-fricative VCVs results with 82.20%, 77.07% and 92.55% mean accuracy for SVM1, SVM2 and HMM, have performed better than other tasks, respectively. Task 4, distinguishing among all 8 consonants with 59.83%, 58.51% and 81.59% for SVM1, SVM2 and HMM, respectively, was the least distinguishable. As the conclusion, in terms of technique, HMM classification technique had better performance over all three studies to classify the consonant segment type of unknown tongue-tip speed profile by the same talker, by a talker from outside the pool and by a talker from within the pool.

Over all three studies, the best classification accuracies were observed for task 1, distinguishing between plosive-fricative VCVs results for SVM1 (90.90%, 82.21% and, 81.47%), SVM2 (89.69%, 77.07% and, 75.07%) and HMM (93.51%, 92.55% and, 92.80%) for study 1, study 2 and, study 3, respectively.

Over all three studies, the worst classification accuracies were observed for Task 4, distinguishing among all 8 consonants for SVM1 (63.24%, 59.83% and, 68.85%), SVM2 (60.57%, 58.51% and, 64.76%) and HMM (82.32%, 81.59% and, 88.30%) for study 1, study 2 and, study 3, respectively.

The difference between the HMM classification accuracies in study one (81.47%) and the HMM classification accuracies in study two (88.3%) among all consonants was found to be significant. Classification of the consonant segment type of unknown tongue-tip trajectories by a talker from within the pool had better results than classification of the consonant segment type of unknown tongue-tip trajectories by a talker from outside the pool.

5.2 Limitations and Future work

This section seeks to describe future work in reference to limitations. In the data collection protocol, each consonant segment is placed with a vowel before and after, to create a vowel-consonant-vowel (VCV) segment. Three carrier vowels were employed: /a/, /i/, /u/, to create a set of 24 different VCVs (3 vowels combined with 8 consonants). In this work, just one of these three carrier vowels, /a/, has been used. Future work should examine the other two carrier vowels (/i/ and /u/) and examine the impact of carrier vowel, if any.

In current work, two different feature sets were employed. Future work should investigate the use of other features for the SVM technique and the use of other parameters for HMM techniques.

All of the research questions of this study refer to *tongue tip kinematic profiles*, which refers to the time-series three-dimensional movement data about the path of the tip of the tongue during the articulation of a speech. Future work should also consider the trajectories of the other active speech articulators as well (e.g., the blade of the the tongue, the lips, and others).

The present study made use of a data set based on 17 talkers. Future work should be based on other speech kinematic datasets, with a larger number of talkers, with a larger number of different speech sounds, and a larger number of repetitions for each speech sound. As well, this dataset contained normative data (talkers without articulatory disorders). Future work should seek to apply and to evaluate classification techniques on kinematic speech signals produced by talkers who have articulatory disorders.

Last, this work did not make use of statistical analyses of the data, which should also be applied in future work. The utility of machine learning techniques over statistical regression analysis should be determined.

Bibliography

Scott G Adams, Gary Weismer, and Raymond D Kent. Speaking rate and speech movement velocity profiles. *Journal of Speech, Language, and Hearing Research*, 36(1):41–54.

Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2014.

Jacqueline Bauman-Waengler. Articulation and phonology in speech sound disorders: A clinical focus 5e. 2016.

Jeffrey J Berry. Accuracy of the NDI WAVE speech research system. *Journal of Speech, Language, and Hearing Research*, 54(5):1295–1301, 2011.

RW Bossemeyer, JG Wilpon, CH Lee, and LR Rabiner. Automatic speech recognition of small vocabularies within the context of unconstrained input. *The Journal of the Acoustical Society of America*, 84(S1):S212–S212, 1988.

Stéphane Canu, Yves Grandvalet, Vincent Guigue, and Alain Rakotomamonjy. SVM and kernel methods matlab toolbox. *Perception Systems et Information, INSA de Rouen, Rouen, France*, 2 (21), 2005.

LJ Cao and WK Chong. Feature extraction in support vector machine: a comparison of pca, xpca and ica. In *Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on*, volume 2, pages 1001–1005. IEEE, 2002.

Winston Chang, Joe Cheng, J Allaire, Yihui Xie, and Jonathan McPherson. Shiny: web application framework for R. *R package version 0.11*, 1, 2015.

O Engwall. Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. 25(1):37–64, 2012.

Arnel C Fajardo and Yoon-joong Kim. Test of vowels in speech recognition using continuous density hidden Markov model and development of phonetically balanced-words in the filipino language. In *Balkan Region Conference on Engineering and Business Education*, volume 1, pages 531–536, 2014.

Susanne Fuchs and Pascal Perrier. *On the complex nature of speech kinematics*. Universitätsbibliothek Johann Christian Senckenberg, 2013.

Susanne Fuchs, Pascal Perrier, Christian Geng, and Christine Mooshammer. What role does the palate play in speech motor control? Insights from tongue kinematics for German alveolar obstruents. *Speech production: Models, phonetic processes, and techniques*, pages 149–164, 2006.

Toni Giorgino. Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of Statistical Software*, 31(7):1–24, 2009.

Colin Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 285–339, 1991.

Frank H Guenther. Skill acquisition, coarticulation, and rate effects in a neural network model of speech production. *The Journal of the Acoustical Society of America*, 95(5):2924–2924, 1994.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

Michael Brandon Haworth. Computer games for motor speech rehabilitation. Master's thesis, York University, Toronto Canada, 2016.

T Hill and P Lewicki. Statistics methods and applications. statssoft, Tulsa, USA, 2007.

Philip Hoole and Noel Nguyen. Electromagnetic articulography. *Coarticulation–Theory, Data and Techniques, Cambridge Studies in Speech Science and Communication*, pages 260–269, 1999.

Md Afzal Hossan, Sheeraz Memon, and Mark A Gregory. A novel approach for MFCC feature extraction. In *Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on*, pages 1–5. IEEE, 2010.

Tokihiko Kaburagi, Kohei Wakamiya, and Masaaki Honda. Three-dimensional electromagnetic articulography: A measurement principle. *The Journal of the Acoustical Society of America*, 118(1):428–443, 2005.

William Katz, Thomas F Campbell, Jun Wang, Eric Farrar, J Coleman Eubanks, Arvind Balasubramanian, Balakrishnan Prabhakaran, and Rob Rennaker. Opti-speech: A real-time, 3D visual feedback system for speech training. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

William F Katz. Influences of electromagnetic articulography sensors on speech produced by healthy adults and individuals with aphasia and apraxia. *Journal of Speech, Language, and Hearing Research*, 49(3):645–659, 2006.

Heejin Kim, Panying Rong, Torrey M Loucks, Mark Hasegawa-Johnson, T Kobayashi, K Hirose, and S Nakamura. Kinematic analysis of tongue movement control in spastic dysarthria. In *INTERSPEECH*, pages 2578–2581, 2010.

Martin Kloster Jensen. Long consonant after short vowel. In *Proceedings of the fourth intenational congress of the phonetic sciences. The Hague: Mouton*, 1968.

Svensson P. Jensen J. Holm T. D. Nielsen M. S. Mosegaard T. fi Baad-Hansen L. Kothari, M. Tongue-controlled computer game: a new approach for rehabilitation of tongue motor function. *Speech production: Models, phonetic processes, and techniques*, 95(3):524–530, 2014.

Christian Kroos. Measurement accuracy in 3D electromagnetic articulography (Carstens AG500). In *Proceedings of the 8th international seminar on speech production*, pages 61–64, 2008.

A. Lofqvist. Tongue movement kinematics in long and short Japanese consonants. *The Journal of the Acoustical Society of America*, 122(1):512–518, 2010.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

Minh Hoai Nguyen and Fernando De la Torre. Optimal feature selection for support vector machines. *Pattern recognition*, 43(3):584–591, 2010.

Joseph S Perkell, Marc H Cohen, Mario A Svirsky, Melanie L Matthies, Iñaki Garabieta, and Michel TT Jackson. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America*, 92(6):3078–3096, 1992.

Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Matthew Richardson, Jeff Bilmes, and Chris Diorio. Hidden-articulator Markov models for speech recognition. *Speech Communication*, 41(2):511–529, 2003.

Krista Rudy. The effect of palate morphology on consonant articulation in healthy speakers. Master's thesis, University of Toronto, Toronto Canada, 2011.

Paul W Schönle, Klaus Gräbe, Peter Wenig, Jörg Höhne, Jörg Schrader, and Bastian Conrad. Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, 31(1):26–35, 1987.

Heidrun Schröter-Morasch and Wolfram Ziegler. Rehabilitation of impaired speech function (dysarthria, dysglossia). *GMS current topics in otorhinolaryngology, head and neck surgery*, 4, 2005.

Helen M Sharp and Stephen M Tasko. Disorders of speech and voice. In *Neurodevelopmental Disabilities*, pages 193–212. Springer, 2011.

Stamp. A revealing introduction to hidden Markov models. 2015.

Fabian Tomaschek, Martijn Wieling, Denis Arnold, and R Harald Baayen. Word frequency, vowel length and vowel quality in speech production: an EMA study of the importance of experience. In *Proceedings of INTERSPEECH*, pages 1302–1306, 2013.

Sarel van Vuuren and Leora R Cherney. A virtual therapist for speech and language therapy. In *International Conference on Intelligent Virtual Agents*, pages 438–448. Springer, 2014.

Jun Wang and Seongjun Hahm. Speaker-independent silent speech recognition with across-speaker articulatory normalization and speaker adaptive training. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Jun Wang, Ashok Samal, Jordan R Green, and Tom D Carrell. Vowel recognition from articulatory position time-series data. In *2009. ICSPCS 2009. 3rd International Conference on Signal Processing and Communication Systems*, pages 1–6. IEEE, 2009.

Jun Wang, Jordan R Green, Ashok Samal, and David B Marx. Quantifying articulatory distinctiveness of vowels. In *Proceedings of INTERSPEECH*, pages 277–280, 2011.

Jun Wang, Ashok Samal, Jordan R Green, and Frank Rudzicz. Whole-word recognition from articulatory movements for silent speech interfaces. *Special Education and Communication Disorders*, (76), 2012.

Jun Wang, Jordan R Green, Ashok Samal, and Yana Yunusova. Articulatory distinctiveness of vowels and consonants: A data-driven approach. *Journal of Speech, Language, and Hearing Research*, 56(5):1539–1551, 2013.

Jun Wang, Ashok Samal, and Jordan Green. Across-speaker articulatory normalization for speaker-independent silent speech recognition. In *Proceedings of INTERSPEECH*, pages 1179–1183, 2014.

John Westbury, Paul Milenkovic, Gary Weismer, and Raymond Kent. X-ray microbeam speech production database. *The of the Acoustical Society of America*, 88(S1):S56–S56, 1990.

Jay G Wilpon, Lawrence R Rabiner, C-H Lee, and ER Goldman. Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(11):1870–1878, 1990.

Yana Yunusova, Jeffrey S Rosenthal, Krista Rudy, Melanie Baljko, and John Daskalogiannakis. Positional targets for lingual consonants defined using electromagnetic articulography.

Yana Yunusova, Jordan R. Green, and Antje Mefferd. Accuracy assessment for AG500, electromagnetic articulograph. *Speech Language, and Hearing Research*, 52:547–555, 2009.

Andreas Zierdt. Problems of electromagnetic position transduction for a three-dimensional articulographic measurement system. *Forschungsberichte-Institut für Phonetik und Sprachliche Kommunikation der Universität München*, 31:137–141, 1993.